npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02372-4

Early diagnosis of axial spondyloarthritis in primary care using multi-agent systems

Xiaojian Ji[1,11], Zhuofeng Li[2,11], Lulu Zeng[1,3,11], Lidong Hu[1], Yanyan Wang[3], Kui Zhang[4], Lianjie Shi[5], Meng Wei[6], Lifeng Chen[7], Lin Guo[8], Jing Dong[9], An'an Wang[9], Lei Sun[9], Yimin Song[9], Huatao Wang[9], Jingming Wang[9], Ying Lei[9], Wenqian Yue[9], Zheng Zhao[1], Jian Zhu[1], Feng Huang[1], Jing Zhang[2] ✉, Tao Li[9,10] ✉ & Kunpeng Li[1] ✉

Axial spondyloarthritis (axSpA) is an inflammatory disease marked by chronic low back pain, with a global average diagnostic delay of 6.7 years. Early diagnosis is crucial for improving prognosis and reducing disability rates, yet primary care physicians (PCPs) may find it challenging to ensure timely recognition and referrals. This study developed and validated Spondyloarthritis Agents (SpAgents), an early diagnostic system based on a multi-agent framework integrating large language models (LLMs) and imaging models. The SpAgents framework includes PlannerAgent, DataAgent, ToolAgent, and DoctorAgent, supported by long-term memory for dynamic knowledge updates. We enrolled 596 patients, dividing 545 from one hospital into a training dataset (n = 359) and a validation dataset (n = 186), along with an independent cohort of 51 patients from five additional hospitals for testing. SpAgents demonstrated strong diagnostic performance, achieving sensitivity of 0.8615 and specificity of 0.8000 during validation, and 0.9375 and 0.7368 during testing. SpAgents exhibited significantly higher sensitivity (0.9400) and accuracy (0.8600) than both PCPs and junior rheumatologists, with overall performance equivalent to that of senior rheumatologists. Under SpAgents-assisted diagnosis, both PCPs and junior rheumatologists showed marked improvements in sensitivity and accuracy. SpAgents effectively enhance early axSpA identification among PCPs, offering an innovative solution to reduce diagnostic delays.

Axial spondyloarthritis (axSpA) is an inflammatory disease primarily characterized by chronic low back pain[1], with a global prevalence ranging from 0.13 to 1.40%[2,3]. The disease typically originates from sacroiliac joint inflammation and progresses to irreversible structural damage, including spinal osteophyte formation and even bamboo spine[4]. Epidemiological studies[5] indicate that 45% of untreated axSpA patients develop disability within 3 years, escalating to 70% by 5 years, severely compromising quality of life. Early diagnosis and timely intervention are critical to delaying disease progression and structural damage, thereby improving prognosis and reducing disability rates. However, diagnostic delays for axSpA remain a global challenge, with an average delay of 6.7 years[6]. A key contributor to this delay lies in primary care: primary care physicians (PCPs) often demonstrate insufficient diagnostic sensitivity for axSpA[7,8]. Evidence highlights significant gaps in PCPs' knowledge, awareness, and confidence regarding axSpA risk factors and hallmark clinical features[9]. Specifically, PCPs exhibit limited recognition of inflammatory back pain (IBP) and other spondyloarthritis (SpA)-related characteristics[10], coupled with inadequate Magnetic Resonance Imaging (MRI) interpretation skills—only 30% of

[1]Department of Rheumatology and Immunology, the First Medical Center, Chinese PLA General Hospital, Beijing, China. [2]Bioinformatics Division, Department of Automation, BNRIST and MOE Key Lab of Bioinformatics, Tsinghua University, Beijing, China. [3]Department of Rheumatology and Immunology, Beijing Electric Power Hospital, Beijing, China. [4]Department of Clinical Immunology, Xijing Hospital, Fourth Military Medical University, Xi'an, China. [5]Department of Rheumatology and Immunology, Peking University Shougang Hospital, Beijing, China. [6]Department of Rheumatology and Immunology, General Hospital of Western Theater Command, Chengdu, China. [7]Department of Rheumatology and Immunology, General Hospital of Central Theater Command, Wuhan, China. [8]Department of Rheumatology and Immunology, General Hospital of Northern Theater Command of Chinese PLA, Shenyang, Liaoning, China. [9]Department of Medical Innovation Research, Chinese PLA General Hospital, Beijing, China. [10]National Engineering Research Center for Medical Big Data Application Technology, Chinese PLA General Hospital, Beijing, China. [11]These authors contributed equally: Xiaojian Ji, Zhuofeng Li, Lulu Zeng. ✉e-mail: 28820327@qq.com; litao30lhospital@163.com; lkp_01@163.com

radiologists are familiar with standardized MRI protocols for sacroiliac joint assessment[11]. These deficiencies prolong diagnostic workflows and increase misdiagnosis risks. Early diagnosis not only improves patient outcomes but also reduces long-term healthcare costs[12,13]. Thus, enhancing PCPs' diagnostic competence for axSpA to facilitate timely referrals represents a pivotal strategy for optimizing disease management.

To enhance early axSpA identification, multiple diagnostic algorithms and tools have been developed to assist PCPs, yet several challenges persist. The Spondyloarthritis Diagnosis Evaluation (SPADE) tool optimizes primary care referrals by integrating 12 clinical features. However, its optimal cutoff threshold remains inconsistent: sensitivity reaches 86% (specificity 49%) at a cutoff score of 2 but plummets to 27% (specificity 92%) when the threshold increases to 3[14]. Additionally, the SPADE scale relies on manual evaluation and lacks automated data extraction, which hinders its integration into clinical workflow and increases the workload of physicians. PCPs widely demand more efficient screening tools compatible with busy clinical workflows to reduce administrative burdens[15]. The UK-developed PRIMIS pop-up system automatically identifies potential axSpA cases within workflows[16]. However, traditional models like PRIMIS are constrained by static training datasets and lack dynamic learning capabilities from real-world diagnostic feedback, which limits their generalizability across diverse populations. Machine learning approaches using electronic health records have been explored for axSpA prediction[17–19]. Notably, Kennedy et al.'s algorithm[18] fails to encompass non-radiographic axSpA (nr-axSpA), diminishing its early warning utility. Furthermore, this model demonstrates low positive predictive value (PPV: 0.15–0.25%) in general populations, rendering it impractical for primary care settings.

This study innovatively proposes a multi-agent collaborative system, which includes four functional submodules: the PlannerAgent responsible for coordinating human-computer interaction and task scheduling, the DataAgent for integrating multimodal data to extract key features, the ToolAgent for invoking MRI image analysis models, and the DoctorAgent for analyzing multidimensional information to generate diagnostic recommendations. Through the collaborative working mechanism of these agents, the system achieves patient data analysis, imaging feature recognition, and medical decision support.

The aim of this study is to construct a multi-agent system for axSpA auxiliary diagnosis that can adapt to real clinical scenarios. By comparing the diagnostic outcomes of real cases between the multi-agent system and physicians of different clinical experience levels, the diagnostic capabilities of the multi-agent system for axSpA are validated. Additionally, the study systematically evaluates the early diagnostic value of the multi-agent system under scenarios with varying clinical data accessibility and the assistance of the ToolAgent, focusing on verifying its diagnostic sensitivity, accuracy, and the effectiveness of medical decision outputs.

Results

Study sample

A total of 596 patients with suspected axSpA were included in this study, comprising 359 in the training set, 186 in the validation dataset, and 51 in the testing dataset. The demographic and clinical characteristics of the study cohorts are summarized in Table 1. Median ages were similar across all cohorts: 31.0 years (25.0–39.0) in the training dataset ($n = 359$), 30.0 years (25.0–38.0) in the validation dataset ($n = 186$), and 29.0 years (25.5–34.0) in the testing dataset ($n = 51$).

Table 1 | Patient characteristics of study samples

Variables	Training Dataset (n = 359)	Validation Dataset (n = 186)	Testing Dataset (n = 51)
Age (years), Med (Q1–Q3)	31.0 (25.0–39.0)	30.0 (25.0–38.0)	29.0 (25.5–34.0)
Male, n(%)	235 (65.5%)	125 (67.2%)	41 (80.4%)
HLA-B27, n(%)	243 (67.7%)	119 (64.0%)	29 (56.9%)
CRP (mg/dL), Med (Q1–Q3)	0.3 (0.1–1.2)	0.3 (0.1–1.0)	0.3 (0.1–0.8)
ESR (mm/h), Med (Q1–Q3)	7.0 (2.0–19.8)	5.0 (2.0–15.0)	3.0 (2.0–8.0)
axSpA, n(%)	263 (73.3%)	134 (72.0%)	33 (64.7%)

Impact of different LLM on SpAgents

This study presents a systematic optimization of LLM configurations within a multi-agent framework, with a particular focus on evaluating the impact of different LLM on diagnostic performance. DeepSeek-Chat was employed uniformly as the core LLM for PlannerAgent, DataAgent and ToolAgent. HuatuoGPT-O1, DeepSeek-Chat, DouBao, Qwen-Plus, and DeepSeek-Reasoner were respectively utilized as the core large language models (LLMs) for the DoctorAgent.

As shown in Table 2, the combination of DeepSeek-Chat and DeepSeek-Reasoner achieves the best overall performance. On the training set, this combination achieved superior performance compared to other configurations, with sensitivity (0.9228), accuracy (0.9083), F1 score (0.9373), and balanced accuracy (0.8947) all outperforming other setups, while specificity (0.8667) was slightly lower. On the test set, it continued to demonstrate a clear advantage, with significantly higher sensitivity (0.8615), accuracy (0.8444), F1 score (0.8889), and balanced accuracy (0.8308), although its specificity (0.8000) remained marginally lower than that of other combinations.

Impact of long-term memory on SpAgents

To assess the contribution of the long-term memory module, we compared SpAgents' diagnostic performance with and without the memory repository, using the optimal DeepSeek-Chat + DeepSeek-Reasoner combination. Both models maintained low UNSURE rates. As shown in Table 3 and Supplementary Fig. S1, in the training dataset, the sensitivity of SpAgents improved from 0.8798 (0.8382–0.9192) to 0.9228 (0.8880–0.9570), specificity from 0.8409 (0.7575–0.9146) to 0.8667 (0.7952–0.9343), and accuracy from 0.8699 (0.8323–0.9017) to 0.9083 (0.8768–0.9398) after integrating the long-term memory. In the validation dataset, sensitivity increased from 0.8134 (0.7500–0.8824) to 0.8615 (0.8000–0.9206) and accuracy from 0.8270 (0.7730–0.8757) to 0.8444 (0.7887–0.8944). In the testing dataset, the sensitivity of SpAgents improved from 0.8750 (0.7500–0.9706) to 0.9375 (0.8485–1.0000), specificity from 0.7368 (0.5294–0.9232) to 0.7368 (0.5263–0.9375), and accuracy from 0.8235 (0.7250–0.9216) to 0.8627 (0.7647–0.9608) after integrating the long-term memory.

Impact of ToolAgent on SpAgents

We compared the diagnostic results of the SpAgents system with and without assistance from the ToolAgent for sacroiliac joint imaging analysis. As shown in Table 4 and Supplementary Fig. S1, the incorporation of the ToolAgent resulted in higher specificity from 0.7717 [0.6829–0.8515] to 0.8667 [0.7975–0.9334] in the training dataset and from 0.7234 [0.5908–0.8421] to 0.8000 [0.6800–0.9057] in the validation dataset.

Performance under varying levels of clinical data availability

This study simulated the progressive availability of clinical data in real-world settings, from initial medical history collection to laboratory and imaging evaluation. We conducts four models (EXP 1 to EXP 4) reflecting increasing levels of data availability, including CRP/ESR, HLA-B27, and MRI report.

The results are presented in Table 5 and Fig. 1. In EXP 1 (only medical history) and EXP 2 (with inflammatory markers), the SpAgents model showed a relatively high "UNSURE rate". As HLA-B27 and MRI data became available in EXP 3 and EXP 4, the "UNSURE rate" dropped significantly to 0.3193 and 0.0367, respectively. EXP 4, which included complete patient data, achieved the highest diagnostic accuracy and the lowest rate of diagnostic uncertainty. Additionally, for patients with incomplete data, SpAgents can extract key information from available clinical data and provide further examination recommendations.

Table 2 | Experimental results of SpAgents performance for different LLMs

Models	Dataset	Sensitivity	Specificity	Accuracy	F1-score
HuatuoGPT-O1	Training Dataset	0.6527	0.8750	0.7123	0.7685
Deepseek-Chat		0.8973	0.7604	0.8607	0.9042
Doubao		0.8429	0.7917	0.8291	0.8782
Qwen-Plus		0.4225	**0.9362**	0.5597	0.5845
DeepSeek-Reasoner		**0.9228**	0.8667	**0.9083**	**0.9373**
HuatuoGPT-O1	ValidationDataset	0.5038	0.8846	0.6108	0.6505
Deepseek-Chat		0.8284	0.7692	0.8118	0.8638
Doubao		0.7970	0.7308	0.7784	0.8379
Qwen-Plus		0.3692	**0.9804**	0.5414	0.5363
DeepSeek-Reasoner		**0.8615**	0.8000	**0.8444**	**0.8889**

The bold values indicates the models with the best performance.

Table 3 | Impact of long-term memory on diagnostic performance

Models	Dataset	UNSURE rate	Sensitivity without UNSURE (95% CI)	Specificitywithout UNSURE (95% CI)	Accuracy without UNSURE (95% CI)	Accuracy with UNSURE (95% CI)	F1-score without UNSURE (95% CI)	Mean Time/s
SpAgents with memory	Training Dataset	0.0279	**0.9228 (0.888–0.9570)**	**0.8667 (0.795–0.9343)**	**0.9083 (0.876–0.9398)**	**0.8830 (0.853–0.9100)**	**0.9373 (0.914–0.9579)**	47.9
SpAgents without memory		0.0362	0.8798 (0.838–0.9192)	0.8409 (0.757–0.9146)	0.8699 (0.832–0.9017)	0.8384 (0.807–0.8700)	0.9098 (0.883–0.9349)	63.5
SpAgents with memory	Validation Dataset	0.0323	**0.8615 (0.800–0.9206)**	0.8000 (0.681–0.9048)	**0.8444 (0.788–0.8944)**	0.8172 (0.756–0.8760)	**0.8889 (0.841–0.9242)**	47.9
SpAgents without memory		0.0053	0.8134 (0.750–0.8824)	**0.8627 (0.758–0.9536)**	0.8270 (0.773–0.8757)	**0.8226 (0.756–0.8760)**	0.8720 (0.825–0.9167)	63.5
SpAgents with memory	Testing Dataset	0.0000	**0.9375 (0.8485–1.0000)**	**0.7368 (0.5263–0.9375)**	**0.8627 (0.7647–0.9608)**	/	**0.8955 (0.8136–0.9677)**	56.3
SpAgents without memory		0.0000	0.8750 (0.7500–0.9706)	0.7368 (0.5294–0.9232)	0.8235 (0.7250–0.9216)	/	0.8615 (0.7619–0.9412)	52.4

The bold values indicates the models with the best performance.

Table 4 | Impact of ToolAgent on diagnostic performance

Models	Dataset	UNSURE rate	Sensitivity without UNSURE (95% CI)	Specificity without UNSURE (95% CI)	Accuracy without UNSURE (95% CI)	Accuracy with UNSURE (95% CI)	F1-score without UNSURE (95% CI)	Mean Time/s
Agents without ToolAgent	Training Dataset	0.0279	**0.9300 (0.8967–0.9575)**	0.7717 (0.6829–0.8515)	0.8883 (0.8567–0.9198)	0.8635 (0.8280–0.898)	0.9246 (0.900–0.9472)	55.8
Agents with ToolAgent		0.0279	0.9228 (0.8884–0.9533)	**0.8667 (0.7975–0.9334)**	**0.9083 (0.8739–0.9398)**	**0.8830 (0.8530–0.9100)**	**0.9373 (0.9149–0.9579)**	47.9
Agents without ToolAgent	Validation Dataset	0.0538	**0.9380 (0.883–0.973)**	0.7234 (0.5908–0.8421)	**0.8807 (0.8295–0.9261)**	**0.8333 (0.7680–0.8920)**	**0.9202 (0.8794–0.9498)**	55.8
Agents with ToolAgent		0.0323	0.8615 (0.7969–0.9174)	**0.8000 (0.6800–0.9057)**	0.8444 (0.7889–0.9000)	0.8172 (0.7560–0.8760)	0.8889 (0.8413–0.9242)	47.9
Agents without ToolAgent	Testing Dataset	0.0000	0.9062 (0.8077–1.0000)	0.7368 (0.5238–0.9287)	0.8431 (0.7451–0.9412)	/	0.8788 (0.7869–0.9552)	54.4
Agents with ToolAgent		0.0000	**0.9375 (0.8485–1.0000)**	**0.7368 (0.5263–0.9375)**	**0.8627 (0.7647–0.9608)**	/	**0.8955 (0.8136–0.9677)**	56.3

The bold values indicates the models with the best performance.

Table 5 | Performance under varying levels of clinical data availability

Models	UNSURE rate	Sensitivity without UNSURE (95% CI)	Specificity without UNSURE (95% CI)	Accuracy without UNSURE (95% CI)	Accuracy with UNSURE (95% CI)	F1-score without UNSURE (95% CI)	Mean time/s
ALL Dataset							
EXP1	0.9009	0.9167 (0.8298–0.9800)	0.1667 (0.0013–0.5000)	0.8333 (0.7222–0.9259)	0.0826 (0.0729–0.0960)	0.9072 (0.8352–0.9608)	65.5
EXP2	0.9101	0.8462 (0.7209–0.9474)	0.7000 (0.3750–0.9219)	0.8163 (0.6939–0.9184)	0.0734 (0.0430–0.1280)	0.8800 (0.7941–0.9487)	50.6
EXP3	0.3193	0.9254 (0.8929–0.9524)	0.5658 (0.4533–0.6774)	0.8518 (0.8328–0.8986)	0.5798 (0.5430–0.6160)	0.9085 (0.8811–0.9316)	47.7
EXP4	0.0367	**0.9326 (0.9079–0.9565)**	**0.7554 (0.6791–0.8231)**	**0.8857 (0.8571–0.9105)**	**0.8532 (0.8240–0.8850)**	**0.9231 (0.9026–0.9424)**	55.8
Training Dataset							
EXP1	0.8942	0.9063 (0.7520–0.9800)	0.1667 (0.0030–0.6410)	0.7895 (0.6620–0.8850)	0.0836 (0.0779–0.0958)	0.8788 (0.7797–0.9577)	65.5
EXP2	0.8914	0.8438 (0.6990–0.9370)	0.5714 (0.2200–0.8660)	0.7948 (0.6620–0.8850)	0.0864 (0.0500–0.1460)	0.8710 (0.7719–0.9538)	50.6
EXP3	0.3008	**0.9343 (0.8940–0.9620)**	0.6415 (0.5260–0.7430)	0.8725 (0.8340–0.9070)	0.6100 (0.5680–0.6500)	0.9204 (0.8906–0.9463)	47.7
EXP4	0.0279	0.9300 (0.8967–0.9575)	**0.7717 (0.6829–0.8515)**	**0.8883 (0.8567–0.9198)**	**0.8635 (0.8280–0.8980)**	**0.9246 (0.9000–0.9472)**	55.8
Validation Dataset							
EXP1	0.9140	0.9375 (0.7300–0.9970)	0.1665 (0.0027–0.6402)	**0.9375 (0.7820–0.9930)**	0.0806 (0.0729–0.0931)	**0.9677 (0.8966–0.9652)**	65.5
EXP2	0.9462	0.8571 (0.4240–0.9970)	0.6000 (0.2920–0.9350)	0.9000 (0.7320–0.9850)	0.0484 (0.0240–0.0579)	0.9231 (0.7273–0.9573)	50.6
EXP3	0.3548	0.9072 (0.8310–0.9560)	0.3913 (0.1980–0.6240)	0.8083 (0.7340–0.8720)	0.5215 (0.4530–0.5890)	0.8844 (0.8360–0.9293)	47.7
EXP4	0.0538	**0.9380 (0.8830–0.9730)**	**0.7234 (0.5908–0.8421)**	0.8807 (0.8295–0.9261)	**0.8333 (0.7680–0.8920)**	0.9202 (0.8794–0.9498)	55.8
Testing Dataset							
EXP1	0.7059	**0.9091 (0.7000–1.0000)**	0.5000 (0.0000–1.0000)	0.8000 (0.5833–1.0000)	0.2353 (0.1373–0.3529)	0.8696 (0.6667–1.0000)	64.2
EXP2	0.6863	0.8000 (0.5000–1.0000)	0.5000 (0.1667–0.8333)	0.6875 (0.5000–0.8750)	0.2157 (0.1209–0.3327)	0.7619 (0.5714–0.9091)	60.4
EXP3	0.2745	0.8400 (0.6800–0.9600)	0.4167 (0.1667–0.6667)	0.7027 (0.5676–0.8378)	0.5098 (0.5676–0.8378)	0.7925 (0.6909–0.8846)	54.7
EXP4	0.0000	0.9062 (0.8077–1.0000)	**0.7368 (0.5238–0.9287)**	**0.8431 (0.7451–0.9412)**	/	**0.8788 (0.7869–0.9552)**	54.4

EXP1: Includes patient medical history records only. EXP2: Incorporates medical records + inflammatory markers (CRP/ESR). EXP3: Adds HLA-B27 testing to the data components of EXP2. EXP4: Further integrates MRI reports into the dataset of EXP3.x
The bold values indicates the models with the best performance.

Analysis of SpAgents and physicians in diagnostic performance

This study assessed the impact of SpAgents on clinical diagnostic performance by comparing physicians' diagnostic metrics before and after using the SpAgents-assisted diagnostic tool (Table 6, Fig. 2, Supplementary Table S1). Compared to physicians, SpAgents achieved significantly higher sensitivity (0.9400 [0.8742–1.0000]) and accuracy (0.8600 [0.7920–0.9280]) than both PCPs and junior rheumatologists (p <0.05), and the overall performance was equivalent to that of senior rheumatologists (å 5 years), and Orthopedist. After a two-week washout period, under SpAgents-assisted diagnosis, PCPs and junior rheumatologists showed marked improvements in sensitivity and accuracy (with sensitivity increasing by 16–34% and accuracy improving by 3–16%, p <0.05). With SpAgents assistance, all primary care physicians demonstrated significant improvements in sensitivity. Doctor 1's sensitivity increased from 0.6800 (0.5507–0.8093) to 0.8400 (0.7384–0.9416) (p = 0.04). Similarly, Doctor 2 showed improvement from 0.6200 (0.4855–0.7545) to 0.8000 (0.6891–0.9109) (p = 0.02), while Doctor 3 exhibited the most substantial gain, with sensitivity rising from 0.4400 (0.3024–0.5776) to 0.7600 (0.6416–0.8784) (p <0.001). After a 3-month washout period, both PCPs and junior rheumatologists maintained significantly improved sensitivity and accuracy with SpAgents assistance compared to baseline, with no statistically significant difference from the 2-week results (p >0.05).

SpAgents diagnostic examples

Fig. 3 presents SpAgents' diagnostic evidence for clinical cases, demonstrating the system's capability to integrate clinical evidence and apply the Assessment of SpondyloArthritis International Society (ASAS) classification criteria. Case 1 (True Positive): A 29-year-old male with persistent back pain for over 1 year. SpAgents integrated clinical symptoms and laboratory findings, then invoked the ToolAgent to identify imaging evidence of active sacroiliitis meeting the ASAS classification criteria. Case 2 (False Negative): A 39-year-old male presenting with back pain consistent with inflammatory back pain. However, his sacroiliac joint inflammation had significantly subsided following treatment and was in an inactive phase. SpAgents utilized the ToolAgent and determined that active sacroiliitis criteria were not met, concluding axSpA criteria were not fulfilled. In addition, the SpAgents system diagnosed "UNSURE" and other examples of cases, see supplementary Table S2.

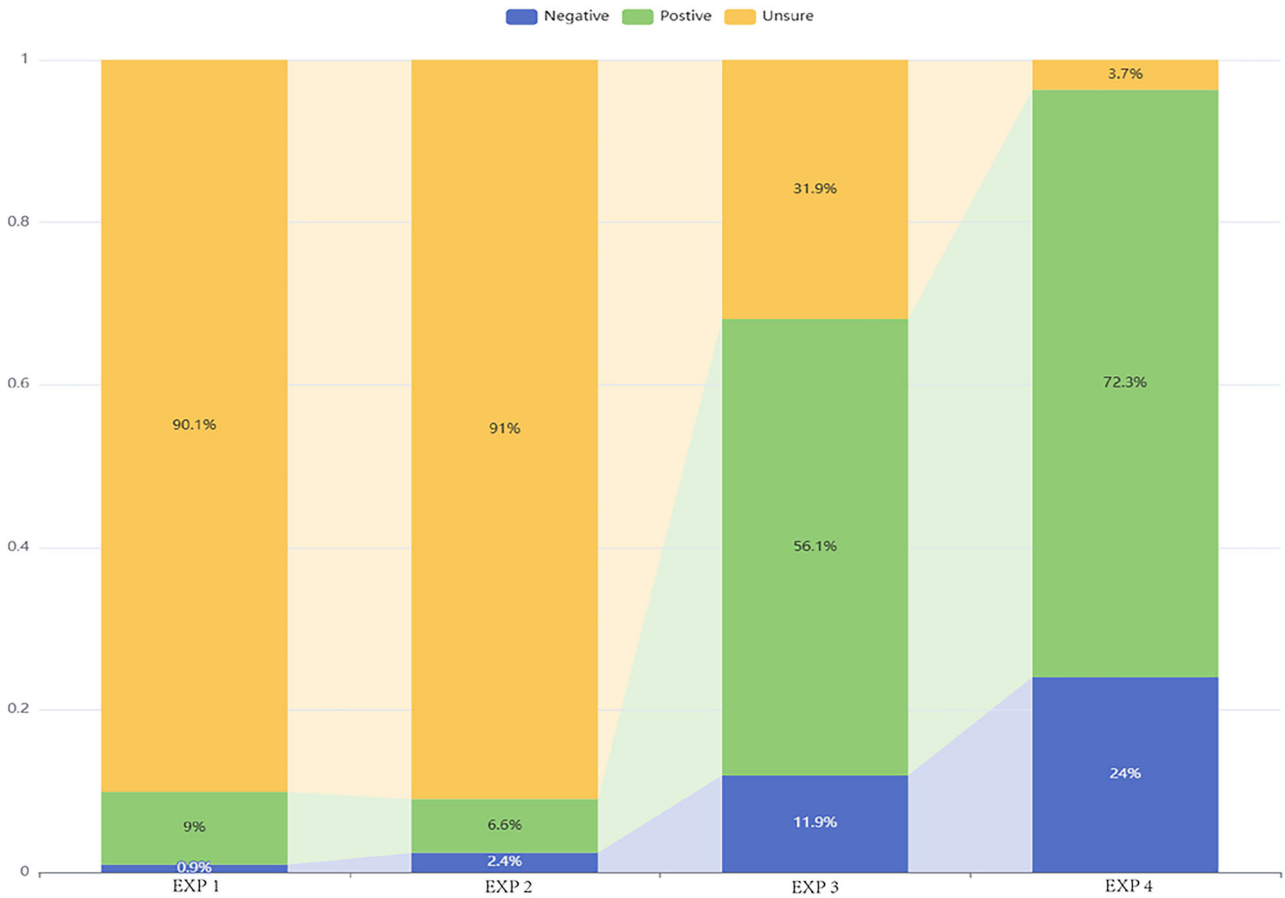


Fig. 1 | Variation in diagnostic outcomes with differing clinical data availability. Analysis of SpAgents and Physicians in Diagnostic Performance. This figure illustrates the variation in diagnostic outcomes of the SpAgents system under different conditions of clinical data availability. The analysis covers four experimental scenarios (EXP 1 to EXP 4), with each scenario incrementally increasing the available clinical data to assess the diagnostic performance of the system. EXP 1 is based solely on the patient's medical history, EXP 2 adds inflammatory markers (such as CRP/ESR), and EXP 3 and EXP 4 introduce HLA-B27 and MRI data, respectively.

Discussion

Principal findings

This study presents SpAgents, a multi-agent system integrating large language and imaging models for multimodal diagnosis of axSpA. SpAgents achieved diagnostic sensitivity and accuracy comparable to senior rheumatologists and outperformed less experienced clinicians, while also enhancing their diagnostic performance. Its memory mechanisms and imaging agents further improved sensitivity and specificity. SpAgents also recognized cases with high diagnostic uncertainty and provided effective decision support, extending its utility in rheumatology practice.

Through systematic evaluation of different LLM combinations, we designed the optimal clinical diagnostic framework. DeepSeek-chat was selected for the PlannerAgent, DataAgent, and ToolAgent due to its superior response speed, while DeepSeek-Reasoner was chosen for the DoctorAgent based on its exceptional core output performance metrics. We demonstrated that the integrated architecture of DeepSeek-Chat and Deepseek-Reasoner achieved optimal clinical diagnostic performance. Notably, the ToolAgent enhanced diagnostic specificity. The long-term memory, which emulates physicians' ability to learn from historical cases, improved the system's recognition of axSpA clinical manifestations. In comparative trials involving seven physicians, SpAgents demonstrated sensitivity and accuracy comparable to senior rheumatologists in axSpA diagnosis, outperforming both PCPs and junior rheumatologists, while exhibiting slightly lower specificity than senior rheumatologists, with consistent performance across both 2-week and 3-month washout periods. Crucially, under SpAgents-assisted mode, both PCPs and junior rheumatologists showed significant improvements in diagnostic performance. Furthermore, the system adaptively provides reliable diagnostic outputs and examination recommendations based on variable clinical data accessibility, addressing critical gaps in existing tools.

Key innovations and clinical advantages

The SpAgents multi-agent framework proposed in this study demonstrates significant performance advantages in the auxiliary diagnosis of axSpA. Unlike existing tools (e.g., SPADE and PRIMIS) that rely on manual feature extraction and static rule-based algorithms, the SpAgents is able to process unstructured, free-text electronic health records in real time. Compared to traditional single imaging models[20] and machine learning models[21], SpAgents innovatively integrates clinical cases, laboratory data, and imaging data in real-time through the DataAgent, which automatically acquires and integrates structured data and unstructured text from electronic health records, extracting potential IBP and other key SpA features. This approach automatically identifies key diagnostic clues and provides timely feedback in a real-world diagnostic environment. Additionally, the context management and exception handling mechanisms introduced by the DataAgent further enhance system stability[22].

Traditional models can only provide diagnostic results but rarely offer explanations for their decisions. In contrast, the DoctorAgent leverages large language model capabilities to simulate the diagnostic reasoning process of physicians. It generates detailed, patient-specific explanations for each diagnosis, thereby enhancing transparency and aiding clinicians in evaluating the tool's recommendations within their own clinical judgment. The DoctorAgent, by constructing a long-term memory, simulates the

Table 6 | Comparative diagnostic performance between SpAgents and physicians across clinical specialties

	Sensitivity(95% CI)	*P* value	Specificity(95% CI)	*P* value	Accuracy(95% CI)	*P* value	F1-score(95% CI)	*P* value
SpAgents	0.9400(0.8742–1.0000)		0.7800(0.6652–0.8848)		0.8600(0.7920–0.9280)		0.8704 (0.8000–0.9358)	
Primary Care Physicians								
Doctor 1 (5 year)								
without SpAgents[a]	0.6800 (0.5507–0.8093)	**<0.001**	0.8000 (0.6891–0.9109)	1.00	0.7400 (0.6540–0.8260)	**0.04**	0.7234 (0.6154–0.8173)	0.19
with SpAgents-2 week	0.8400 (0.7384–0.9416)	**0.04**	0.8400 (0.7384–0.9416)	0.77	0.8400 (0.7681–0.9119)	**0.01**	0.8400 (0.7578–0.9076)	**0.01**
with SpAgents-3 month[b]	0.8800 (0.7959–0.9600)	**<0.001**	0.8200 (0.7111–0.9167)	0.488	0.8500 (0.7800–0.9200)	**0.011**	0.8544 (0.7783–0.9231)	**<0.001**
Doctor 2 (8 years)								
without SpAgents[a]	0.6200 (0.4855–0.7545)	**<0.001**	0.9400 (0.8742–1.0000)	**0.03**	0.7800 (0.6988–0.8612)	0.180	0.7381 (0.6250–0.8409)	**0.01**
with SpAgents-2 week	0.8000 (0.6891–0.9109)	**0.02**	0.8200 (0.7135–0.9265)	0.08	0.8100 (0.7331–0.8869)	**<0.001**	0.8081 (0.7143–0.8866)	0.08
with SpAgents-3 month[b]	0.8200 (0.7154–0.9224)	**0.002**	0.8400 (0.7420–0.9311)	0.092	0.8300 (0.7600–0.9000)	**<0.001**	0.8283 (0.7513–0.8987)	**0.038**
Doctor 3 (15 years)								
without SpAgents[a]	0.4400 (0.3024–0.5776)	**<0.001**	0.9200 (0.8392–0.9813)	0.10	0.7100 (0.6211–0.7989)	**0.01**	0.6027 (0.4615–0.7180)	0.14
with SpAgents-2 week	0.7600 (0.6416–0.8784)	**<0.001**	0.9200 (0.8448–0.9952)	0.25	0.8400 (0.7681–0.9119)	**<0.001**	0.8261 (0.7272–0.9053)	**<0.001**
with SpAgents-3 month[b]	0.8000 (0.6764–0.9111)	**<0.001**	0.9000 (0.8163–0.9787)	0.865	0.8500 (0.7800–0.9152)	**0.002**	0.8421 (0.7558–0.9111)	**0.001**
Rheumatologists								
Doctor 4 (1 year)								
without SpAgents[a]	0.3400 (0.2087–0.4713)	**<0.001**	0.9600 (0.9057–1.0000)	**0.02**	0.6500 (0.5565–0.7435)	**0.002**	0.4928 (0.3379–0.6342)	**0.02**
with SpAgents-2 week	0.6800 (0.5507–0.8093)	**<0.001**	0.9400 (0.8742–1.0000)	1.00	0.8100 (0.7331–0.8869)	**<0.001**	0.7816 (0.6665–0.8696)	**<0.001**
with SpAgents-3 month[b]	0.7400 (0.6092–0.8515)	**<0.001**	0.8800 (0.7778–0.9600)	0.984	0.8100 (0.7300–0.8800)	**<0.001**	0.7957 (0.6966–0.8763)	**<0.001**
Doctor 5 (6 years)								
without SpAgents[a]	0.8600 (0.7638–0.9562)	0.340	0.8600 (0.7638–0.9562)	0.290	0.8600 (0.7920–0.9280)	0.810	0.8603 (0.7875–0.9263)	0.91
with SpAgents-2 week	0.9200 (0.8448–0.9952)	0.51	0.9000 (0.8168–0.9832)	0.48	0.9100 (0.8539–0.9661)	**<0.001**	0.9035 (0.7936–0.9301)	0.08
with SpAgents-3 month[b]	0.9400 (0.8696–1.0000)	0.149	0.8800 (0.7778–0.9619)	0.517	0.9100 (0.8548–0.9600)	0.03	0.9126 (0.8526–0.9636)	0.065
Doctor 6 (12 years)								
without SpAgents[a]	0.8800 (0.7899–0.9701)	0.50	0.9000 (0.8168–0.9832)	0.08	0.8900 (0.8287–0.9513)	0.63	0.8889 (0.8139–0.9444)	0.97
with SpAgents-2 week	0.9400 (0.8742–1.0000)	0.25	0.9200 (0.8448–0.9952)	1.00	0.9300 (0.8800–0.9800)	**<0.001**	0.9307 (0.8764–0.9787)	**0.02**
with SpAgents-3 month[b]	0.9200 (0.8247–0.9923)	0.315	0.9400 (0.8645–1.0000)	0.736	0.9300 (0.8700–0.9800)	0.028	0.9293 (0.8601–0.9775)	0.044
Orthopedist								
Doctor 7 (10 years)								
without SpAgents[a]	0.8600 (0.7638–0.9562)	0.29	0.7400 (0.6184–0.8616)	0.75	0.8000 (0.7216–0.8784)	0.24	0.8113 (0.7200–0.8908)	0.34
with SpAgents-2 week	0.9800 (0.9412–1.0000)	0.08	0.7000 (0.5730–0.8270)	0.48	0.8400 (0.7681–0.9119)	**<0.001**	0.8596 (0.7916–0.9219)	0.06
with SpAgents-3 month[b]	0.9800 (0.9332–1.0000)	**0.04**	0.7200 (0.5957–0.8470)	0.778	0.8500 (0.7800–0.9200)	**<0.001**	0.8673 (0.7943–0.9355)	0.052

Doctor 1, Doctor 2 and Doctor 3 are from the general medicine department, with clinical experience of 5 year, 8 years, and 15 years, respectively. Doctor 4, Doctor 5 and Doctor 6 are from the rheumatology department, with clinical experience of 1 year, 6 years, and 12 years, respectively. Doctor 7 is from the orthopedics department, with 10 years of clinical experience. § The *P* value for the comparison between SpAgents and doctors without SpAgents. ¶ The *P* value for the comparison between doctors without SpAgents and doctors with SpAgents after a 2-week washout period. # The *P* value for the comparison between doctors without SpAgents and doctors with SpAgents after a 3-month clearance period. Boldface values indicate the *P* values with statistically significant differences.

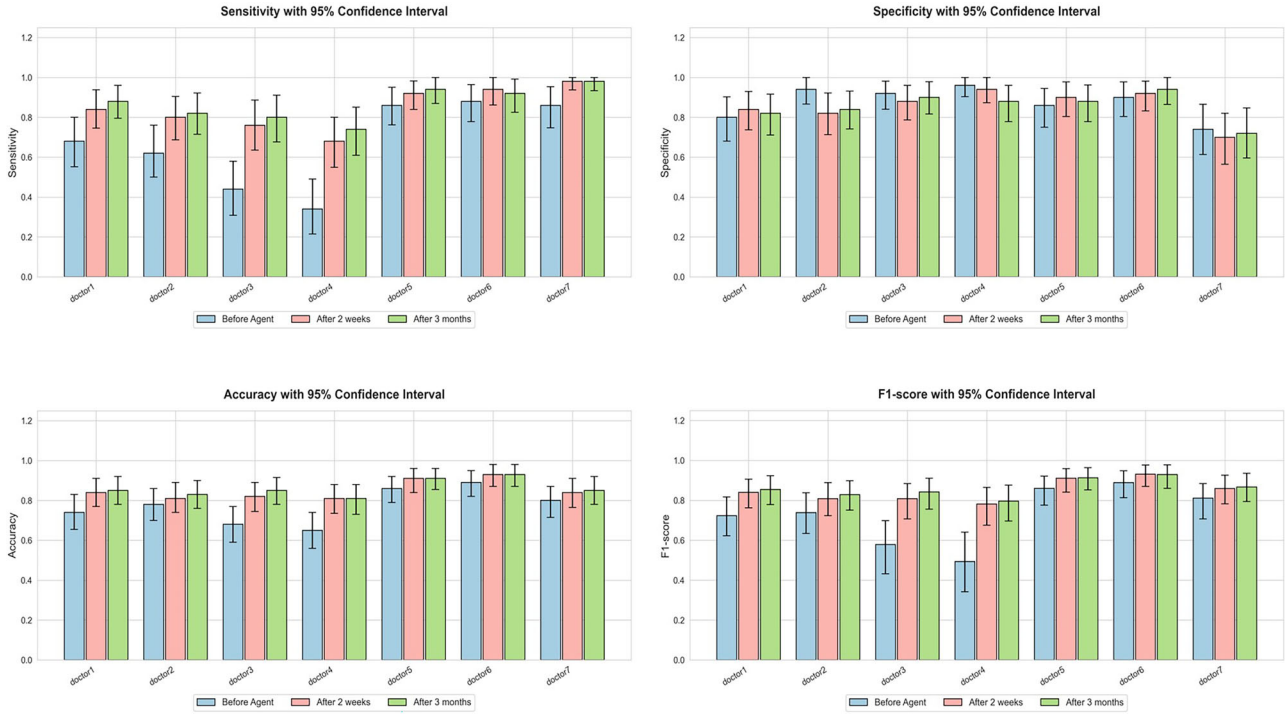


Fig. 2 | Diagnostic performance metrics of SpAgents versus physicians across clinical specialties. This figure illustrates the diagnostic performance metrics (sensitivity, specificity, accuracy, and F1-score) of primary care physicians and specialists in diagnosing axSpA. The metrics are displayed before and after the use of SpAgents. This visualization highlights the enhancement in diagnostic capabilities among individual physicians with varying levels of clinical experience, demonstrating the positive impact of the SpAgents system on diagnostic outcomes.

mechanism by which physicians continuously learn and reference past cases during clinical diagnosis. The introduction of the long-term memory enables the agents to dynamically learn and update knowledge based on physician feedback during the auxiliary diagnostic process. As demonstrated in our case analyses, referencing historical cases allows the system to improve its reasoning and diagnostic accuracy over time.

The ToolAgent developed in this study improved diagnostic specificity while maintaining sensitivity and accuracy by invoking an external axSpA-specialized imaging model to analyze patient MRI data. This enhancement holds critical clinical significance, effectively reducing the risk of misdiagnosing bone marrow edema findings as axSpA[23]. Unlike traditional standalone imaging analysis methods[24,25], SpAgents' multi-agent architecture enables synergistic reasoning between imaging features and clinical data, overcoming the limitations of prior machine learning models that focused solely on sacroiliac joint imaging. The diagnostic reliability of axSpA heavily depends on precise MRI interpretation of sacroiliac joints. However, the complex anatomical structure of these joints and the high expertise threshold for imaging interpretation often lead to diagnostic errors among less-experienced physicians[26]. SpAgents addresses this gap through agent collaboration, integrating multidimensional clinical data to provide reliable imaging analysis support for primary care physicians, thereby reducing their experience-related limitations. The specificity improvement achieved by the ToolAgent demonstrates the system's advantage in minimizing false-positive outcomes. Notably, our multi-agent framework features a modular design that provides extensibility for future integration of additional imaging tools, such as deep learning models developed by Bordner et al.[27] and Lee et al.[28].

This study systematically evaluated the diagnostic performance of the SpAgents across varying data availability scenarios by simulating clinical decision-making pathways. The framework was validated through a stepwise clinical workflow from initial consultation to progressively incorporating laboratory tests and imaging assessments. Results demonstrated a stepwise improvement in diagnostic certainty and accuracy with incremental clinical data. In the primary evaluation phase, the model automatically identified key SpA features such as IBP and family history of SpA. Among 54 patients flagged as axSpA-positive based on multiple characteristic features, 44 were ultimately confirmed with axSpA. While prior studies developed referral tools based on symptomatic presentations[29], their limited accuracy and sensitivity hinder practical utility. Our SpAgents demonstrates superior performance, particularly valuable for regions with constrained clinical resources. Notably, the integration of inflammatory markers (CRP/ESR) provided limited diagnostic improvement, consistent with existing evidence: only 25–40% of axSpA patients exhibit elevated CRP[30,31], and normal CRP/ESR cannot exclude diagnosis, while elevated levels may reflect non-axSpA etiologies[32]. The "UNSURE rate" decreased progressively with additional tests, most markedly after incorporating HLA-B27 testing and MRI. HLA-B27, the strongest independent genetic risk factor, and MRI-detected sacroiliitis provided critical diagnostic evidence[32], underscoring their necessity in evaluating suspected axSpA. Unlike conventional algorithms requiring complete datasets[33,34], our framework uniquely adapts to real-world clinical workflows. By outputting "UNSURE" when data are insufficient, it maintains diagnostic reliability while minimizing false negatives. This reasoning mode aligns with clinical logic: deferring definitive diagnosis until adequate evidence is available. Thus, SpAgents serves dual roles: as a triage tool for primary care to prioritize referrals and as a decision-support system in hospitals to enhance diagnostic precision. Importantly, the average computational cost per diagnosis was approximately 0.0161 CNY (~0.0023 USD) (see Supplementary Table S3 for detailed token statistics), highlighting the cost-effectiveness of this approach for clinical deployment.

This study has several limitations. First, while the current ToolAgent can identify and quantify bone marrow edema from a single MRI sequence, it cannot distinguish inflammatory from non-inflammatory edema (such as fractures or infections) or identify structural lesions and spinal changes. Second, the system currently relies on a file-based data source. While this

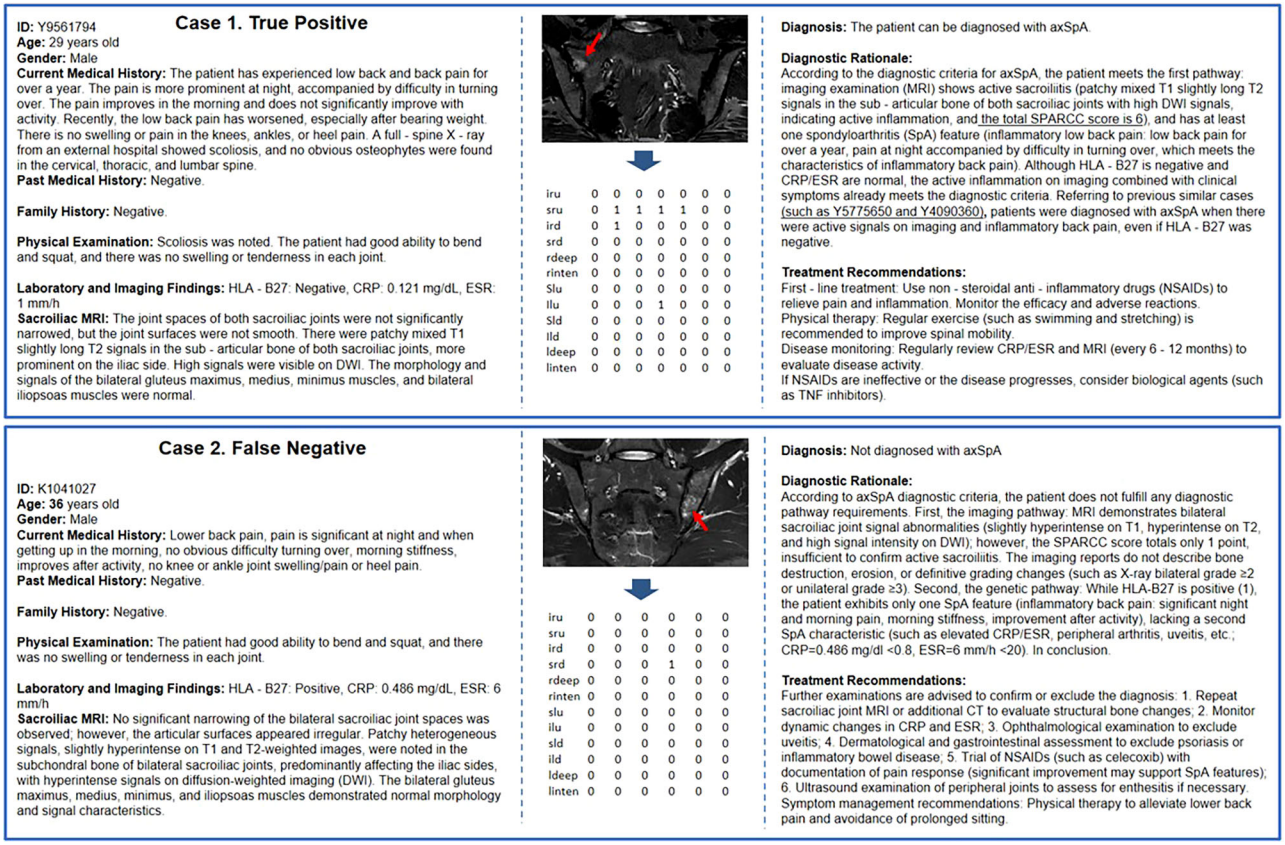


Fig. 3 | Demonstrates the application of SpAgents in the diagnosis of axSpA. This figure illustrates the diagnostic evidence from SpAgents for clinical cases. Case 1 (True Positive): A 29-year-old male with persistent back pain for over a year, where SpAgents successfully identified axSpA through integrated clinical symptoms, laboratory findings, and imaging evidence. Case 2 (False Negative): A 39-year-old male with back pain indicative of inflammatory back pain, but with inactive sacroiliac joint inflammation post-treatment; SpAgents determined that the criteria for axSpA were not met. RUI right upper ilium, RLI right lower ilium, RUS right upper sacrum, RLS right lower sacrum, LUI left upper ilium, LLI left lower ilium, LUS left upper sacrum, LLS left lower sacrum, SPARCC Spondyloarthritis Research Consortium of Canada, axSpA Axial Spondyloarthritis, CRP C-reactive Protein, ESR Erythrocyte Sedimentation Rate, HLA-B27 Human Leukocyte Antigen-B27.

approach enabled focused development and validation of the core algorithms and workflow, integration with hospital information systems is crucial for broader clinical applicability. Only with such integration can SpAgents be thoroughly validated and evaluated in real-world clinical settings. Finally, the inherent "black-box" nature of LLMs poses challenges for the transparency and traceability of medical decision-making. To address this, we have made explicit explanation output a core requirement in the DoctorAgent's design, aiming to enhance the system's interpretability and user trust.

Prospects for scalability and enhancement of SpAgents

SpAgents features a modular architecture that supports flexible adaptation to diverse clinical workflows and data environments. Its core components, including the PlannerAgent, DataAgent, and ToolAgent, can be readily customized through prompt template adjustments, terminology mapping, and tailored API integration to meet the unique requirements of various healthcare institutions. The DoctorAgent's compatibility with multiple large language models further allows for seamless adaptation to differing computational resources and technological infrastructures.

SpAgents can be further optimized through ToolAgent model enhancements. Future enhancements will expand ToolAgent's functionality beyond bone marrow edema detection to include recognition of bone erosion and fat infiltration via multi-sequence MRI integration, enabling more comprehensive identification of axSpA subtypes. Additionally, the SpAgents design supports future implementation of adjustable ToolAgent probability thresholds to optimize sensitivity-specificity balance, particularly important for minimizing missed diagnoses in clinical practice.

Key development priorities include deep integration with hospital information systems to automate data flow and advance from simulated environments to real-world deployment. Extensive clinical validation across large-scale, prospective, multi-center studies will address selection bias and strengthen international generalizability. Complemented by user interface optimization, standardized training protocols, and rigorous Electronic Medical Record/Picture Archiving and Communication System integration testing, these efforts will establish SpAgents as a robust clinical decision support tool capable of transforming diagnostic workflows in diverse healthcare settings.

While the current model demonstrates strong performance on multi-center data, its generalization capability in broader, diverse prospective patient populations remains to be validated. Our planned approaches include: (1) We will further expand the dataset to rigorously evaluate SpAgents' diagnostic performance in real-world prospective clinical settings. (2) Since the current system's imaging features are limited to bone marrow edema, we will develop and integrate modules for other sacroiliac joint lesions, thereby enhancing the system's performance. (3) We will conduct large-scale multicenter clinical trials to evaluate SpAgents' integration capabilities with different hospital information systems, operational stability in real-world workflows, and its impact on improving diagnostic efficiency and confidence among physicians at different levels.

In summary, SpAgents, as a multi-agent framework based on LLMs and imaging models, demonstrates excellent performance in the auxiliary diagnosis of axSpA. It significantly enhances the diagnostic efficacy of physicians, particularly in primary care settings, highlighting its important clinical value.

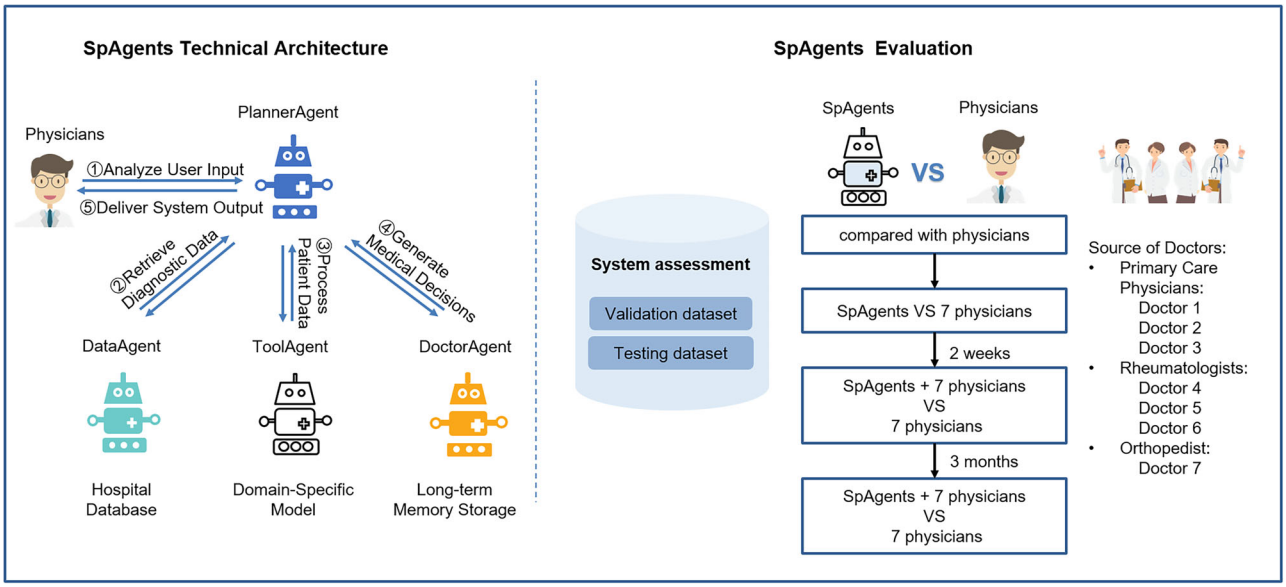
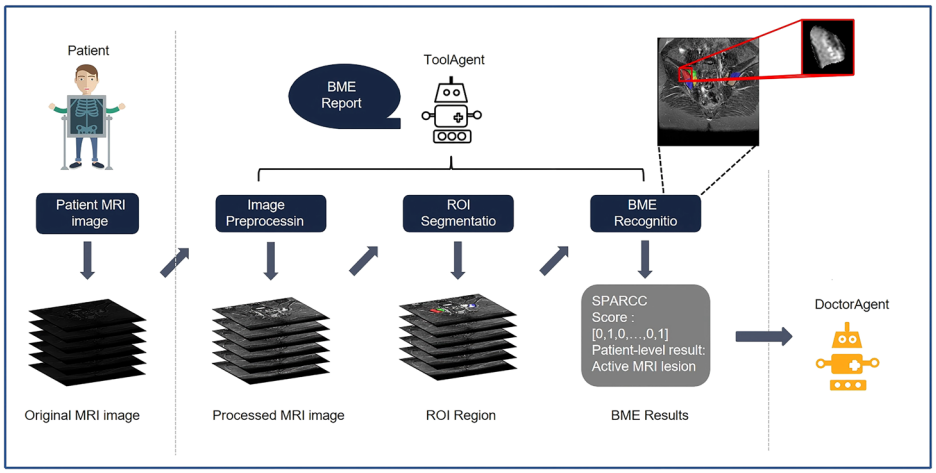


Fig. 4 | Overview of the study. This figure outlines the workflow of the SpAgents multi-agent system. On the left, it illustrates the functions of various agents: the PlannerAgent manages tasks, the DataAgent integrates diverse patient data, the ToolAgent performs MRI analysis, and the DoctorAgent synthesizes diagnostic decisions. On the right, the figure compares SpAgents-assisted diagnostics with human evaluations.

Fig. 5 | Illustration of ToolAgent workflow for automated SPARCC scoring. This figure illustrates the operational workflow of the ToolAgent within the SpAgents system, designed to enhance the diagnosis of axSpA through automated SPARCC scoring. The process begins with image preprocessing of the patient's MRI data, followed by segmentation of the region of interest within the sacroiliac joints using a 3D U-Net model. Subsequently, it performs bone marrow edema classification and quantification, generating comprehensive SPARCC scores based on the processed images.



Methods

Study design

This study proposes a multi-agent system, SpAgents, for the diagnosis of axSpA, based on LLMs and imaging models, Supplementary Fig. S2. SpAgents consists of four submodules: PlannerAgent, DataAgent, ToolAgent, and DoctorAgent. The overall research process is shown in Fig. 4. This system categorizes the tasks of four agents into two distinct types, with tailored prompt designs based on their characteristics. The first category comprises step-by-step simple tasks (such as planning, retrieval, and invocation) executed by PlannerAgent, DataAgent, and ToolAgent. We ensure standardized operations by defining specific execution flow constraints to prevent behavioral divergence. The second category involves complex medical reasoning and decision-making independently handled by DoctorAgent. In terms of their specific functions: DataAgent retrieves patient clinical data from databases and handles file uploads; ToolAgent integrates with external deep learning models to process MRI images through two-stage segmentation and generate Spondyloarthritis Research Consortium of Canada (SPARCC) scores across 12 anatomical regions, presenting the results in tabular format for input to DoctorAgent; PlannerAgent orchestrates the workflow by determining which agents to invoke based on user queries; and DoctorAgent synthesizes processed data from both DataAgent and ToolAgent to make final diagnostic decisions. Our prompts strictly adhere to the ASAS-EULAR axSpA classification standard, ensuring rigorous and logically sound diagnostic processes. The algorithmic workflow and prompt design for SpAgents' invocation process are detailed in Supplementary Note 1.Algorithm and Supplementary Note 2.Prompt.

(1) PlannerAgent interacts with physicians using natural language. Then PlannerAgent interprets diagnostic intents of users, decomposes them into subtasks, and coordinates specialized agents to execute corresponding tasks. To ensure robust performance, PlannerAgent integrates a context management module for maintaining conversational coherence and an exception handling framework for resolving invocation errors.

(2) DataAgent retrieves and integrates patient's information of various modalities. The patient's information includes textual clinical records

(such as symptoms and family history), laboratory test results, DICOM-format MRI data, and corresponding imaging reports.

(3) ToolAgent functions to invoke external tools for processing multimodal data, including edema quantification scoring models that analyze patient MRI scans. Specifically, it invokes specialized models for sacroiliac joint bone marrow edema recognition, performing both qualitative diagnosis and quantitative assessment using the one-stop SPARCC scoring system, as shown in Fig. 5.

(4) DoctorAgent gives the final diagnostic result including classifications of axSpA, non-axSpA, or UNSURE. Based on physicians' diagnostic feedback, the DoctorAgent can update its long-term memory to store clinical cases for future reference, thereby progressively enhancing diagnostic accuracy.

This study systematically optimizes configuration schemes for LLMs within a multi-agent framework. The framework comprises four agents categorized into two types based on their tasks: The first type (including PlannerAgent, DataAgent, and ToolAgent) handles procedural tasks such as step planning, information retrieval, and tool invocation. These tasks require high response speed but relatively low complex reasoning capabilities. Therefore, we uniformly selected the DeepSeek-Chat model for these three agents. The second type (DoctorAgent) performs core medical reasoning and final diagnostic decision-making tasks, whose outputs directly determine the system's diagnostic performance. To evaluate the performance of different large language models in this critical role, we compared multiple models including Huatuo GPT-O1, DeepSeek-Chat, Douban, Qwen-Plus, and DeepSeek-Reasoner., and selected the optimal configuration. The model's utility in primary care settings was assessed by comparing its performance to that of PCPs and specialists with varying levels of experience. In addition, we evaluated the impact of ToolAgent on diagnostic performance, the diagnostic capabilities under varying levels of clinical data availability, and the role of long-term memory in enhancing the diagnostic capacity of SpAgents.

Data preparation

This study included patients who visited the Department of Rheumatology and Immunology at the First Medical Center of the Chinese People's Liberation Army General Hospital between January 2011 and October 2023 due to low back pain. The inclusion criteria were: (1) presence of low back pain; (2) age ≥18 years; (3) completion of sacroiliac joint MRI examination with imaging data including oblique coronal T1-weighted sequences and T2 fat-suppressed sequences; (4) availability of Human leukocyte antigen-B27 (HLA-B27) test results. The exclusion criteria were: (1) pregnant women; (2) presence of tumor or infectious diseases; (3) missing original DICOM files of MRI. In total, 545 patients (397 with axSpA, 148 with non-axSpA) were included and divided into a training dataset ($n = 359$) and a validation dataset ($n = 186$), following the data partitioning protocol established during the ToolAgent imaging model development to prevent potential data leakage. To further conduct external validation, a testing dataset ($n = 51$) included patients from six different medical institutions: Beijing Electric Power Hospital (18 cases), General Hospital of Western Theater Command (11 cases), Xijing Hospital (4 cases), Peking University Shougang Hospital (8 cases), General Hospital of Northern Theater Command of Chinese PLA (2cases) and General Hospital of Central Theater Command (8 cases).

The final diagnosis for all patients was determined based on outpatient follow-up and comprehensive clinical data by two rheumatologists with over 10 years of clinical experience. For cases with diagnostic discrepancies, a third senior rheumatology expert was consulted, and a consensus diagnosis was established after discussion. This study was approved by the Ethics Committee of the Chinese People's Liberation Army General Hospital (Approval No. S2022-255-01). All procedures involving patient data strictly adhered to relevant ethical guidelines and regulations. All retrospective patient data used were collected/processed in strict accordance with the procedures approved by the ethics committee. Individual informed consent for the use of retrospective data was waived due to the nature of the study and the extent of data anonymization or de-identification. All patient data were anonymized, including the deletion of DICOM header metadata.

All patient data were stored in the MySQL 5.7 database, anonymized, and transmitted via SSL encryption to ensure compliance with the Personal Information Protection Law. The research network was based on the hospital's internal LAN, and data access was strictly managed using a role-based access control (RBAC) system.

Design of ToolAgent

ToolAgent can utilize medical image recognition models to enhance axSpA diagnosis. We have integrated a trained SPARCC scoring model into ToolAgent. This model automatically generates SPARCC scores based on the patient's fat-suppressed sequence images, providing quantitative assessment of bone marrow edema.The process includes: (1) images preprocessing using the SimpleITK (version 2.3.1); (2) region of interest (ROI) segmenting using 3D U-Net to isolate quadrant-level sacroiliac joint regions; (3) bone marrow edema classification using ResNet architecture; and (4) compilation of classification results into a structure SPARCC score report. As shown in Fig. 5.

Design of DoctorAgent

The diagnostic accuracy of DoctorAgent directly impacts the overall diagnostic performance of SpAgents, and its algorithm is detailed in the Supplementary Note 1.Algorithm. During system initialization, the DoctorAgent initializes key components such as the long-term memory repository and the LLM based on configuration files. Subsequently, the DoctorAgent may receive either a "diagnosis" or "learning" task.

Upon receiving a "diagnosis" task, DoctorAgent first searches for the top k (where $k = 3$ in this study) most similar cases in the long-term memory based on the patient's input information. Specifically, it uses ClinicalBERT[35] to encode the patient data into a query vector and employs the FAISS library to compute cosine similarity between the query vector and the stored case vectors. The top k most similar vectors are decoded into corresponding case texts. These retrieved cases, including both patient details and diagnostic outcomes, are returned to DoctorAgent. Then, DoctorAgent integrates the current patient's information, retrieved reference cases, prior medical knowledge of axSpA, and output requirements into a prompt, which is then fed into the LLM. To address cases with incomplete patient information, the system includes a fallback option: "Insufficient information, unable to diagnose." This results in three possible diagnostic outcomes including "Diagnosed as axSpA", "Diagnosed as non-axSpA" and "Diagnosis uncertain (UNSURE)". This output logic mimics real-world clinical reasoning, where physicians typically recommend further tests rather than give a definitive diagnosis when patient's data is inadequate for a definitive diagnosis. For cases diagnosed as axSpA, the system outputs a diagnostic rationale along with treatment recommendations. For cases diagnosed as non-axSpA or UNSURE, it provides reasons and suggestions for further medical examination.

If the task type is "learning", the agent dynamically updates the long-term memory repository based on diagnostic feedback from physicians to continuously improve diagnostic performance. Specifically, the agent uses ClinicalBERT[35] to encode the patient data and diagnostic label into a 768-dimensional vector and stores it in the memory for future reference. Notably, the agent immediately updates the memory repository upon receiving feedback. For retrieval operations, the memory repository implements vector search functionality through the FAISS [1.10.0] library. Additionally, the Long-Term Memory Repository provides data management capabilities, supporting CRUD operations (Create, Read, Update, Delete) on stored data to ensure dynamic updates and maintainability.

Evaluation of diagnostic model

The reasoning capability of the LLM in DoctorAgent directly influences the diagnostic performance of SpAgents. To obtain the most effective SpAgents, we sequentially used HuatuoGPT-O1[36], DeepSeek-Chat[37], DouBao[38],

Qwen-Plus[39], and DeepSeek-Reasoner[40] as the core LLMs for DoctorAgent and evaluated each model's performance using key metrics such as sensitivity, specificity, accuracy, F1-score, and balanced accuracy. Model outputs labeled as "UNSURE" were excluded from the metric calculations. To provide a conservative estimate of diagnostic performance, all "UNSURE" outputs were classified as incorrect predictions when calculating the "accuracy with UNSURE" metric.

Evaluation of the long-term memory

The long-term memory is designed to simulate the way physicians leverage past case experience. The update and retrieval processes are illustrated in Supplementary Fig. S3. We conducted comparative experiments with and without the long-term memory repository to evaluate its effect on diagnostic performance. Metrics including sensitivity, specificity, accuracy, F1-score, and balanced accuracy were calculated for comparison.

Evaluation of ToolAgent for imaging analysis

ToolAgent currently integrates a tool for qualitative and quantitative analysis of sacroiliac joint bone marrow edema. This tool consists of four modules: image preprocessing, ROI segmentation, edema identification (including qualitative evaluation of active sacroiliitis and SPARCC scoring), and LLM-based report generation. We evaluated the diagnostic performance of SpAgents with and without the ToolAgent.

Evaluation under varying levels of clinical data availability

To simulate real-world clinical scenarios, we designed a progressive data accessibility framework to evaluate SpAgents' adaptability in real diagnostic environments. Initially, only patient demographic characteristics (sex, age) and textual medical records such as chief complaints, history of present illness, and past medical history are provided. Subsequently, key clinical data are incrementally incorporated, including C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), HLA-B27, and textual reports of sacroiliac joint MRI imaging. SpAgents make decisions based on different data information respectively and provide corresponding diagnostic results to assist clinical decision-making.

Evaluation of SpAgents in assisting physicians diagnosis

To compare SpAgents with physicians, we implemented random sampling to select 50 patients from each group (axSpA/non-axSpA) from a total of 545 patients across the training dataset and validation dataset, yielding a balanced cohort of 100 patient cases for physicians vs. SpAgents comparison. Seven licensed physicians from four medical institutions participated in an evaluation: three PCPs (with 5, 8, and 15 years of clinical experience), three rheumatologists (with 1, 6, and 12 years of clinical experience), and one orthopedic surgeon (10 years of clinical experience).

A two-stage comparative experimental design was implemented. In the first stage, physician independently evaluated diagnose anonymized patients without prior knowledge of individual diagnoses or the overall prevalence ratio between patients with and without axSpA. A 2-week and 3-month washout period was established to minimize potential learning effects and evaluation bias. In the second phase, (SpAgents-assisted phase), the same physicians re-diagnosed the same case set (patient cases randomly reordered) with decision support from the SpAgents system at 2 weeks[41,42] and 3 months post-randomization. Diagnostic performance metrics including sensitivity, specificity, accuracy, F1-score were systematically calculated for each physician under both unassisted and SpAgents-assisted conditions.

Statistical analysis

All statistical analyses were conducted using Python version 3.12.10. key packages included pandas (version 2.2.2) for data handling, NumPy (version 1.26.4) for numerical computation, scipy (version 1.13.1) for McNemar's test and confidence interval estimation. Bootstrap resampling for F1-score confidence intervals was implemented manually using NumPy. Confidence intervals for model performance metrics (including sensitivity, specificity, accuracy, and F1-score) were estimated using a nonparametric bootstrap resampling method with 1000 iterations. This approach generated empirical distributions through random patient-case sampling with replacement, with 95% percentile-based confidence intervals derived from these distributions. Statistical differences between each clinician and the AI system were assessed using McNemar's test for paired nominal data on accuracy, sensitivity, and specificity. For F1-score comparisons, bootstrap-based difference testing was employed to directly evaluate score differences through resampling. A $p < 0.05$ was considered statistically significant.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to institutional confidentiality policies and patient privacy regulations, but are available from the corresponding author on reasonable request.

Code availability

Code is openly available at the following link for non-commercial purpose: https://github.com/SpAgents/SpAgents.

Received: 31 July 2025; Accepted: 13 January 2026;
Published online: 22 January 2026

References

1. Navarro-Compán, V., Sepriano, A., El-Zorkany, B. & van der Heijde, D. Axial spondyloarthritis. *Ann. Rheum. Dis.* **80**, 1511–1521 (2021).
2. Stolwijk, C., van Onna, M., Boonen, A. & van Tubergen, A. Global prevalence of spondyloarthritis: a systematic review and meta-regression analysis. *Arthritis Care Res.* **68**, 1320–1331 (2016).
3. Bohn, R., Cooney, M., Deodhar, A., Curtis, J. R. & Golembesky, A. Incidence and prevalence of axial spondyloarthritis: methodologic challenges and gaps in the literature. *Clin. Exp. Rheumatol.* **36**, 263–274 (2018).
4. Schwartzman, S. & Ruderman, E. M. A road map of the axial spondyloarthritis continuum. *Mayo Clin. Proc.* **97**, 134–145 (2022).
5. Zhao, Y. et al. Current health resources required for the management of ankylosing spondylitis in developing areas of China. *Chin. Med. J.* **136**, 737–739 (2023).
6. Zhao, S. S. et al. Diagnostic delay in axial spondyloarthritis: a systematic review and meta-analysis. *Rheumatology* **60**, 1620–1628 (2021).
7. Garrido-Cumbrera, M. et al. Identifying parameters associated with delayed diagnosis in axial spondyloarthritis: Data from the European map of axial spondyloarthritis. *Rheumatology* **61**, 705–712 (2022).
8. Gaffney, K., Webb, D. & Sengupta, R. Delayed diagnosis in axial spondyloarthritis-How can we do better?. *Rheumatology* **60**, 4951–4952 (2021).
9. Steen, E., McCrum, C. & Cairns, M. Physiotherapists' awareness, knowledge and confidence in screening and referral of suspected axial spondyloarthritis: a survey of UK clinical practice. *Musculoskelet. Care* **19**, 306–318 (2021).
10. Coath, F. L. & Gaffney, K. Inflammatory back pain: a concept, not a diagnosis. *Curr. Opin. Rheumatol.* **33**, 319–325 (2021).
11. Barnett, R., Gaffney, K. & Sengupta, R. Diagnostic delay in axial spondylarthritis: a lost battle?. *Best. Pract. Res Clin. Rheumatol.* **37**, 101870 (2023).
12. Berbel-Arcobé, L. et al. Association between diagnostic delay and economic and clinical burden in axial spondyloarthritis: a multicentre retrospective observational study. *Rheumatol. Ther.* **12**, 255–266 (2025).
13. Yi, E., Ahuja, A., Rajput, T., George, A. T. & Park, Y. Clinical, economic, and humanistic burden associated with delayed diagnosis of axial spondyloarthritis: a systematic review. *Rheumatol. Ther.* **7**, 65–87 (2020).
14. Habibi, S., Doshi, S. & Sengupta, R. THU0413 utility of the spade tool to identify axial spondyloarthritis in patients with chronic backpain. *Ann. Rheum. Dis.* **75**, 338–338 (2016).

15. Lapane, K. L. et al. Primary care physician perspectives on screening for axial spondyloarthritis: A qualitative study. *PLoS ONE* **16**, e0252018 (2021).
16. PRIMIS, ttps://www.nottingham.ac.uk/primis/projects/axspa.aspx
17. Sengupta, R. et al. P261 Early and accurate diagnosis of patients with axial spondyloarthritis using machine learning: a predictive analysis from electronic health records in the United Kingdom. *Rheumatology* **73**, 4017–4018 (2022).
18. Kennedy, J. et al. Predicting a diagnosis of ankylosing spondylitis using primary care health records-A machine learning approach. *PLoS ONE* **18**, e0279076 (2023).
19. Walsh, J. A., Rozycki, M., Yi, E. & Park, Y. Application of machine learning in the diagnosis of axial spondyloarthritis. *Curr. Opin. Rheumatol.* **31**, 362–367 (2019).
20. Adams, L. C., Bressem, K. K. & Poddubnyy, D. Artificial intelligence and machine learning in axial spondyloarthritis. *Curr. Opin. Rheumatol.* **36**, 267–273 (2024).
21. Redeker, I. et al. Identification of a machine learning-based diagnostic model for axial spondyloarthritis in rheumatological routine care using a random forest approach. *RMD Open* **10**, e004702 (2024).
22. Abbasian, M., Azimi, I., Rahmani, A. M. & Jain, R. Conversational health agents: a personalized LLM-powered agent framework. *JAMIA Open* **8**, ooaf067 (2025).
23. Seven, S. et al. Anatomic distribution of sacroiliac joint lesions on magnetic resonance imaging in patients with axial spondyloarthritis and control subjects: a prospective cross-sectional study, including postpartum women, patients with disc herniation, cleaning staff, runners, and healthy individuals. *Arthritis Care Res.* **73**, 742–754 (2021).
24. Lee, S. et al. Artificial intelligence for the detection of sacroiliitis on magnetic resonance imaging in patients with axial spondyloarthritis. *Front. Immunol.* **14**, 1278247 (2023).
25. Lin, K. Y. Y., Peng, C., Lee, K. H., Chan, S. C. W. & Chung, H. Y. Deep learning algorithms for magnetic resonance imaging of inflammatory sacroiliitis in axial spondyloarthritis. *Rheumatology* **61**, 4198–4206 (2022).
26. Diekhoff, T. & Ziegeler, K. Anatomical variation of the sacroiliac joints - what the rheumatologist should know. *Curr. Opin. Rheumatol.* https://doi.org/10.1097/bor.0000000000001091.(2025).
27. Bordner, A. et al. A deep learning model for the diagnosis of sacroiliitis according to assessment of SpondyloArthritis International Society classification criteria with magnetic resonance imaging. *Diagn. Inter. Imaging* **104**, 373–383 (2023).
28. Lee, K. H., Choi, S. T., Lee, G. Y., Ha, Y. J. & Choi, S. I. Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning. *Diagnostics* **11**, https://doi.org/10.3390/diagnostics11071156.(2021).
29. Braun, A. et al. Optimizing the identification of patients with axial spondyloarthritis in primary care-the case for a two-step strategy combining the most relevant clinical items with HLA B27. *Rheumatology* **52**, 1418–1424 (2013).
30. Rudwaleit, M. et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. *Arthritis Rheum.* **60**, 717–727 (2009).
31. van den Berg, R. et al. Percentage of patients with spondyloarthritis in patients referred because of chronic back pain and performance of classification criteria: experience from the Spondyloarthritis Caught Early (SPACE) cohort. *Rheumatology* **52**, 1492–1499 (2013).
32. van Gaalen, F. A. & Rudwaleit, M. Challenges in the diagnosis of axial spondyloarthritis. *Best. Pr. Res. Clin. Rheumatol.* **37**, 101871 (2023).
33. Zhang, K. et al. Use of MRI-based deep learning radiomics to diagnose sacroiliitis related to axial spondyloarthritis. *Eur. J. Radio.* **172**, 111347 (2024).
34. Jia, W. et al. Ankylosing spondylitis prediction using fuzzy K-nearest neighbor classifier assisted by modified JAYA optimizer. *Comput Biol. Med.* **175**, 108440 (2024).
35. Alsentzer, E. et al. Publicly available clinical BERT embeddings. ClinicalNLP (2019).
36. Chen, J. et al. HuatuoGPT-o1, towards medical complex reasoning with LLMs. In *Findings of the Association for Computational Linguistics* (2025).
37. Liu, A. et al. DeepSeek-V3 technical report. https://ui.adsabs.harvard.edu/abs/2024arXiv241219437D (2024).
38. DouBao, <https://www.doubao.com/chat/.
39. Qwen. et al. Qwen2.5 technical report. https://ui.adsabs.harvard.edu/abs/2024arXiv241215115Q (2024).
40. Guo, D. et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. https://ui.adsabs.harvard.edu/abs/2025arXiv250112948D (2025).
41. Fu, H. et al. AI assisted reader evaluation in acute CT head interpretation (AI-REACT): protocol for a multireader multicase study. *BMJ Open* **14**, e079824 (2024).
42. Lee, S. J. et al. Using a Deep Learning-Based Decision Support System to Predict Emergent Large Vessel Occlusion Using Non-Contrast Computed Tomography. *J Clin Med* **14**, https://doi.org/10.3390/jcm14134635.(2025).

Acknowledgements

We sincerely thank all physicians who participated in our human-computer comparative clinical trials for their valuable expertise and contributions. This work was supported by Beijing Natural Science Foundation (L242143), National Key Research and Development Program of China (2021ZD0140409), Youth Independent Innovation Science Fund Project of Chinese PLA General Hospital (22QNFC139).

Author contributions

X.J., Z.L., and L.Z.: Conceptualization, Methodology, Writing–Original Draft. Y.W., K.Z., L.S., M.W., L.C., and L.G.: Data Acquisition, Manuscript Revision. J.D., A.W., L.S., and Y.S.: Investigation, Resources. H.W., J.W., Y.L., W.Y., and L.H.: Software, Formal Analysis. Z.Z., J.Z., and F.H.: Supervision. K.L., T.L., and J.Z.: Conceptualization, Funding Acquisition, Writing–Review & Editing.

Competing interests

The authors declare no competing interests.

Declaration of Generative AI in Scientific Writing

We confirm that no generative AI tools (such as ChatGPT or other large language models) were used in any portion of the manuscript generation. All content in this manuscript was independently prepared by the authors

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02372-4.

Correspondence and requests for materials should be addressed to Jing Zhang, Tao Li or Kunpeng Li.