Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An agentic AI workflow for spine history taking and surgeon-facing clinical synthesis

Begüm Aslantaş Kaplan

begumaslantas@gmail.com

Tokat Erbaa State Hospital

Ali Kaplan

Tokat Erbaa State Hospital

Ali Aydilek

Gülhane Training and Research Hospital

Özkan Çeliker

Denizli Private Saglik Hospital

Tolga Ege

University of Health Sciences, Gülhane Training and Research Hospital

Salim Şentürk

Üsküdar University

İlker Solmaz

Private Practice

Article

Keywords: agentic AI, large language models, spine, history-taking, clinical decision support, digital health

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10617202/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Adequate history-taking is central to safe spine triage but is often constrained by outpatient time pressure. We developed SpineGPT-2, an agentic history-taking system combining a patient-facing interview agent, a deterministic orchestration layer enforcing safety and history coverage, and a surgeon-facing agent that produces a structured, retrieval-grounded summary. In a prospective single-clinic study, 240 patients were allocated by alternating clinic days to the model arm (n = 120) or standard consultation (n = 120). An independent expert panel rated summary adequacy on a five-item scale. Mean adequacy was 4.23 (95% CI, 4.10–4.36), significantly exceeding the pre-specified actionable threshold of 4.0. The panel's leading diagnosis fell within the model's top-three differentials in 88.3% of cases; transcript-based red-flag sensitivity and specificity were 93.8% and 96.2%. Physician-led history-taking time fell from 10.22 to 6.77 minutes. Hallucination occurred in 8.3% of summaries. Agentic, safety-first history-taking produced clinically adequate surgeon summaries, supporting supervised pre-consultation use.

Introduction

The diagnostic power of the medical history has been well-documented in general outpatient medicine. Peterson et al. reported that history-taking led to the final diagnosis in 61 of 80 patients, and Sandler similarly showed that the history was the most important contributor to diagnostic decision-making in a large medical outpatient cohort[1–2]. This principle is also relevant to patients with spinal complaints, because the initial history can distinguish axial mechanical pain from radicular pain, neurogenic claudication, inflammatory or systemic pain patterns, and presentations suggestive of serious spinal pathology[3–7]. A targeted anamnesis can also identify features such as previous malignancy, unintentional weight loss, history of major trauma, prolonged corticosteroid use, urinary or bowel dysfunction, and rapidly progressive neurological symptoms, which may redirect the diagnostic pathway toward conditions that require urgent care. Therefore, history-taking in spine care practice should be viewed not merely as a preliminary conversation but as a high value diagnostic and triage instrument that structures clinical reasoning by converting a broad complaint such as back or neck pain into a clinically meaningful diagnostic pathway, identifies high-risk presentations, prioritizes differential diagnoses, and determines the rational use of physical examination and imaging[3–9].

While the the literature does not define an optimal duration for spine history-taking, it supports the concept that adequate time for a structured spine history is essential for diagnostic safety, surgical triage, avoidance of unnecessary imaging, and rational allocation of specialist resources[10–16]. On the other hand, spine-specific workload studies suggest that the demand for specialist assessment often exceeds available physician capacity, creating an environment in which comprehensive history-taking may be constrained by referral volume and long waiting times[12–16]. These studies suggest that the main challenge is not the absence of knowledge about how to take a spine history, but the difficulty of preserving enough clinical time for history-taking in systems burdened by high patient volume, limited

specialist availability, and non-clinical workload; which can ultimately yielding incomplete histories and occasionally, delayed recognition of clinically important warning features[12–16]. Therefore, a system that could reliably conduct a structured, complete spine history before the patient encounter and hand the surgeon an actionable synthesis would therefore address a consequential clinical need.

Large language models (LLMs) encode substantial clinical knowledge and demonstrated impressive performance across a range of medical knowledge tasks, including licensing-style examinations and structured clinical question-answering benchmarks[17–18]. However, most published evaluations remain static, single-turn, or vignette-based; meaning that the model is asked to respond after the relevant clinical information has already been assembled[19–20]. This is different from real clinical practice, where diagnostic reasoning often begins before the case is fully formed and depends on the ability to elicit, refine, and prioritize information through history-taking. Prompt-engineering studies have shown that a standalone LLM, when appropriately prompted and constrained, can function as a clinically useful interviewer: it can ask focused follow-up questions, maintain a differential diagnostic frame, and organize patient-reported information into a coherent history[21–23]. The underlying language model is, in this sense, already capable of serving as the conversational and reasoning engine of clinical anamnesis. Yet a standalone LLM remains limited as a clinical deployment unit, because it is not, by itself, equipped with the surrounding infrastructure needed to place that interview within a controlled workflow.

Agentic AI addresses this gap by embedding the LLM within an operational architecture. In such systems, the LLM remains as the cognitive engine, but it is connected to external capabilities that may include workflow state, memory, retrieval, tool use, structured outputs, or predefined action policies[24]. Figure 1 illustrates the conceptual shift from a standalone prompt–response language model to an agentic system in which the LLM operates as the cognitive engine within a goal-directed architecture that maintains task state, accesses external resources, applies action policies, and generates structured outputs for downstream use.

Two broad properties distinguish agentic systems from standalone LLM use: autonomous operation and goal-directed behaviour[25]. Rather than simply responding to isolated prompts, an agent can pursue a defined task over a sequence of actions and adjust its next steps according to the current state of that task, which makes it far better suited to tasks that unfold over many turns. Together, these properties shift the system from a passive responder into an active participant in a workflow. This is the capability profile that the task of patient history-taking actually requires, and several recent systems have been built and evaluated along these lines[26–28]. Nevertheless, much of this evidence comes from simulated or general-care settings, leaving open the question of whether an agentic history-taking system can perform safely and usefully with real patients in a specialty-specific consultation.

We developed SpineGPT-2 to address this question in spine outpatient care. The system builds on our earlier prompt-engineering–based SpineGPT framework, which showed that a standalone LLM can elicit a clinically organised spine history when given task-specific instructions and clinical constraints[23]. SpineGPT-2 extends that foundation by embedding the history-taking task within an agentic architecture

composed of three serial components: a patient-facing agent, a rule-based orchestration layer, and a surgeon-facing agent. This design shifts the evaluation from whether an LLM can conduct a structured spine interview under prompting conditions to whether an agentic system can generate a surgeon-facing summary during routine outpatient care. The primary outcome is the clinical adequacy of this summary the main output of SpineGPT-2, rated by an independent expert panel. Pre-specified secondary and exploratory outcomes address diagnostic concordance and accuracy, red-flag detection, consultation-time efficiency, content accuracy and hallucination, operational safety events, information-gap taxonomy localising where the interview most often under-collected information, and patient-reported acceptability of the agent interaction. We hypothesised that, in routine outpatient practice, the agentic system produces a clinical history synthesis that meets this pre-specified standard of adequacy, evidence that would move LLM-based spine anamnesis from demonstrated feasibility toward genuine clinical utility.

Results

Participants and baseline characteristics

Over 30 clinic days during the enrolment period, 240 consecutive patients attending the spine outpatient clinic of a single surgeon were allocated by an alternating day-block design to the agent-based history-taking system (model arm, n = 120) or to classic clinician history-taking (control arm, n = 120). The two arms were enrolled on 15 and 15 clinic days respectively. All 240 patients were retained in the analysis and analysed according to their allocated study arm. Incomplete model-arm encounters were retained in the analysis and are reported as operational safety events.

Mean age was 50.7 ± 13.0 years in the model arm and 50.8 ± 12.3 years in the control arm. Degenerative and mechanical conditions were the most common diagnostic categories, accounting for 99 of 120 patients in the model arm and 101 of 120 patients in the control arm. Serious spinal pathologies were also present in the cohort, including osteoporotic vertebral compression fracture, spinal infection, neoplastic disease, and cauda equina syndrome; these diagnoses accounted for 16 patients in the model arm and 15 patients in the control arm. An objective reference diagnosis was available for 85 patients in the model arm (70.8%) and 80 patients in the control arm (66.7%), based on imaging, laboratory testing, or another clinically accepted diagnostic investigation. Baseline characteristics are summarised in Table 1, and the diagnosis-level distribution by study arm is provided in Supplementary Table S3.

Table 1
Baseline characteristics by study arm

Characteristic	Model arm (n = 120)	Control arm (n = 120)
Age, years, mean (SD)	50.7 (13.0)	50.8 (12.3)
Sex, n (%)		
Female	74 (61.7)	66 (55.0)
Male	46 (38.3)	54 (45.0)
Predominant symptomatic region, n (%)		
Cervical	20 (16.7)	22 (18.3)
Thoracic	10 (8.3)	10 (8.3)
Lumbar	89 (74.2)	87 (72.5)
Multiple	1 (0.8)	1 (0.8)
Diagnostic category, n (%)		
Degenerative / mechanical	99 (82.5)	101 (84.2)
Inflammatory axial spondyloarthritis	5 (4.2)	4 (3.3)
Serious spinal pathology	16 (13.3)	15 (12.5)
Objective reference diagnosis available, n (%)	85 (70.8)	80 (66.7)

Values are mean (SD) or n (%). Percentages are column percentages within each study arm. Serious spinal pathology included osteoporotic vertebral compression fracture, spinal infection, neoplastic disease, and cauda equina syndrome. An objective reference diagnosis was defined as a diagnosis confirmed by imaging, laboratory testing, or another clinically accepted diagnostic investigation. SD, standard deviation.

Primary outcome: Adequacy of the surgeon-facing summary

The primary endpoint was the adequacy of the SFA summary, rated by an independent expert panel using a five-item, five-point scale (1 = strongly disagree to 5 = strongly agree). Because the panel evaluated only model-generated summaries against an absolute standard, raters were aware that the summaries were system-generated. To reduce anchoring by the model's own conclusions, panellists first recorded their differential diagnoses and safety-critical assessments from the interview transcript before the SFA's Diagnostic Impression and Recommended Next Steps sections were revealed.

Across the 120 model-arm encounters, the mean adequacy score was 4.23 (SD 0.71; 95% CI, 4.10–4.36). A one-sample, one-sided t-test against the pre-specified safely actionable threshold of 4.0 showed that

mean adequacy significantly exceeded this threshold ($t = 3.54$, $df = 119$, $p < 0.001$; Cohen's $d = 0.32$). Inter-rater reliability for the overall adequacy score was high, with an average-measures, two-way random-effects, absolute-agreement ICC of 0.88 (95% CI, 0.83–0.91). Overall, 76.7% of summaries achieved a mean adequacy score of at least 4, indicating that in most consultations the panel judged the model-generated history sufficient to support confident clinical use. Item-level performance was broadly consistent across the rubric. The highest mean scores were observed for determining the next diagnostic steps (Q2, mean 4.35) and identifying safety-critical features (Q4, mean 4.34). The lowest-scoring item was whether the history was sufficient to proceed without repeating the full clinician history (Q5, mean 3.92; 71.7% scoring ≥ 4), reflecting the more demanding nature of this criterion and the residual need for targeted surgeon questioning in some cases. The distribution of case-level adequacy scores and the item-level estimates are illustrated in Fig. 2, with exact numerical results provided in Supplementary Table S4.

Secondary and exploratory outcomes

Diagnostic concordance. Diagnostic concordance between the model's ranked differential diagnosis and the expert-panel consensus was assessed descriptively. The panel's consensus primary diagnosis appeared within the model's top-three differential diagnoses in 106 of 120 cases (88.3%; 95% CI, 81.4–92.9) and exactly matched the model's first-ranked diagnosis in 87 cases (72.5%; 95% CI, 63.9–79.7For reference, all three panellists selected the same primary diagnosis in 93 of 120 cases (77.5%), and inter-rater agreement for the categorical primary-diagnosis assignments was high (Fleiss' $\kappa = 0.821$). These findings provide a contextual benchmark for interpreting model–panel diagnostic concordance. In the pre-specified sub-analysis restricted to the 85 patients with an objective reference diagnosis, the confirmed diagnosis appeared within the model's top-three differential diagnoses in 73 cases (85.9%; 95% CI, 76.9–91.7) and was ranked first in 56 cases (65.9%; 95% CI, 55.3–75.1). Diagnostic concordance estimates are shown in Fig. 3a, with detailed denominators and Wilson 95% confidence intervals provided in Supplementary Table S5.

Red-flag detection. Red-flag detection by the SFA was benchmarked against the panel's transcript-based red-flag judgement. Of the 16 encounters flagged by the panel from the interview transcript, the model reported a red flag in 15 and missed 1, corresponding to a sensitivity of 93.8% (95% CI, 71.7–98.9). Among the 104 transcripts judged red-flag-negative by the panel, the model agreed in 100 cases and generated 4 false-positive alerts, corresponding to a specificity of 96.2% (95% CI, 90.5–98.5). These estimates were interpreted descriptively given the low prevalence of true red flags. As an exploratory secondary analysis, the SFA red-flag report was also compared with the treating surgeon's post-examination verification. Against this reference, the model identified 14 of 17 verified red flags and missed 3, yielding a sensitivity of 82.4% (95% CI, 59.0–93.8), while specificity was 95.1% (98 of 103; 95% CI, 89.1–97.9). The corresponding confusion matrices and performance estimates are shown in Fig. 3b, with full contingency counts provided in Supplementary Table S6.

Consultation-time efficiency. Physician-led history-taking time was significantly shorter in the model arm than in the control arm. The mean duration was 6.77 ± 1.61 minutes in the model arm and 10.22 ± 2.46 minutes in the control arm, corresponding to an average reduction of 3.45 minutes. Median history-taking time was 6.50 minutes (IQR, 5.70–7.78) with SpineGPT-2 support and 10.00 minutes (IQR, 8.20–11.53) under standard consultation. In patient-level Mann–Whitney U analysis, this difference was statistically significant ($U = 1579.5$, $P < 0.001$), with a large rank-biserial effect size ($r = -0.78$). The distribution of physician-led history-taking times by study arm is shown in Fig. 4, with detailed descriptive statistics provided in Supplementary Table S7.

Content accuracy and hallucination. A senior author assessed content accuracy across the full model arm by comparing the system-log interview transcript with the SFA output side by side. Hallucination, defined as information stated in the summary but not supported by the transcript, was identified in 10 of 120 cases (8.3%), comprising 10 individual instances: 4 fabricated symptoms, 3 fabricated history elements, 2 temporal errors, and 1 misattribution. Unsupported clinical inference was identified in 6 cases (5.0%), and an inaccurate retrieval-provenance label was identified in 8 cases (6.7%). No hallucination involved a safety-critical red-flag item. Detailed distributional summaries are provided in Supplementary Table S8.

Operational safety incidents. Safety surveillance covered the full model arm of 120 patients across three pre-defined incident categories. Technical failure occurred in 5 encounters, comprising 2 connectivity drops and 3 session freezes. Patient-initiated early termination occurred in 2 encounters, and patient-facing hard-rule violation, defined as a diagnostic assertion or treatment recommendation reaching the patient in the chat interface, was identified in 2 encounters. In total, 9 operational safety events were recorded. Event-category counts, rates, and descriptions are provided in Supplementary Table S9.

Failure taxonomy: Information gaps. When the expert panel judged that clinically relevant information was missing, each gap was mapped to one of eight predefined clinical domains. A total of 62 information gaps were recorded across 33 encounters, mainly in summaries with lower adequacy scores. The most frequent omissions concerned prior treatment and response[14], functional impact[11], prior diagnostic workup[9], pain trajectory or course[8], past medical history or comorbidities[7], systemic red-flag screening[5], neurological inquiry[5], and pain characterisation[3]. The complete distribution of information-gap instances across the eight predefined domains is shown in Supplementary Table S10.

Patient experience. Acceptability of the agent interaction was assessed in the model arm using the Chatbot Usability Questionnaire (CUQ; 16 items, standardised to a 0–100 score). No patient-experience instrument was administered in the control arm. Across the 120 model-arm patients, the mean CUQ score was 74.6 (SD 11.8; range, 44–96). This places the patient-facing interaction in an acceptable usability range and supports the feasibility of using the system as a pre-consultation interview tool. Full descriptive statistics for the CUQ are provided in Supplementary Table S11.

Secondary and exploratory outcomes are summarised in Table 2.

Table 2
Summary of secondary and exploratory outcomes

Outcome	Estimate	Interpretation / denominator
Diagnostic concordance		
Top-1 diagnostic concordance vs expert-panel consensus	87/120 (72.5%) (95% CI, 63.9–79.7)	Model first-ranked diagnosis exactly matched the panel consensus primary diagnosis.
Top-3 diagnostic concordance vs expert-panel consensus	106/120 (88.3%) (95% CI, 81.4–92.9)	Panel consensus primary diagnosis appeared anywhere in the model top-three differential.
Panel unanimity for primary diagnosis	93/120 (77.5%) (95% CI, 69.2–84.1)	All three panellists assigned the same primary diagnosis.
Fleiss kappa for panel primary diagnosis	0.821	Computed across the three independent panel raters.
Confirmed diagnosis ranked first by the model	56/85 (65.9%) (95% CI, 55.3–75.1)	Restricted to cases with objective reference diagnosis.
Confirmed diagnosis within model top-three differential	73/85 (85.9%) (95% CI, 76.9–91.7)	Restricted to cases with objective reference diagnosis.
Red-flag detection		
Sensitivity vs expert-panel transcript judgement	15/16 (93.8%) (95% CI, 71.7–98.9)	TP = 15, FN = 1.
Specificity vs expert-panel transcript judgement	100/104 (96.2%) (95% CI, 90.5–98.5)	TN = 100, FP = 4.
Sensitivity vs treating-surgeon post-examination verification	14/17 (82.4%) (95% CI, 59.0–93.8)	TP = 14, FN = 3.
Specificity vs treating-surgeon post-examination verification	98/103 (95.1%) (95% CI, 89.1–97.9)	TN = 98, FP = 5.
Consultation-time efficiency		
Physician-led history-taking time, model arm	6.50 min (IQR, 5.70–7.78); mean 6.77 (SD 1.61)	Pre-visit SpineGPT-2 interview time excluded.
Physician-led history-taking time, control arm	10.00 min (IQR, 8.20–11.53); mean 10.22 (SD 2.46)	Standard surgeon history-taking interval.
Mann–Whitney U test	U = 1579.5, P < 0.001	Two-sided asymptotic $P = 1.44 \times 10^{-25}$
Rank-biserial effect size	r = − 0.78	Negative sign indicates shorter times in the model arm under the coding used.

Outcome	Estimate	Interpretation / denominator
Diagnostic concordance		
Content accuracy and provenance		
Hallucinated content	10/120 (8.3%)	Clinically relevant information in summary not supported by transcript.
Unsupported clinical inference	6/120 (5.0%)	Diagnostic/management inference not justified by recorded history or retrieved evidence.
Retrieval-provenance label error	8/120 (6.7%)	Citation or provenance label did not correspond to retrieved session evidence.
Operational safety and reliability		
Technical failure	5/120 (4.2%)	Connectivity drop or session freeze.
Patient-initiated early exit	2/120 (1.7%)	Patient stopped before structured closure.
Patient-facing hard-rule violation	2/120 (1.7%)	Diagnostic assertion or treatment recommendation reached the patient.
Information gaps and patient experience		
Encounters with at least one information gap	33/120 (27.5%)	Total gap instances = 62.
CUQ usability score		
Model arm	74.6 (SD 11.8; range 44–96)	Chatbot Usability Questionnaire, standardized 0–100 score.

Diagnostic concordance, red-flag detection, content accuracy, operational safety events, information gaps, and CUQ were assessed in the model arm unless otherwise specified; consultation-time efficiency compared the model and control arms. Confidence intervals for proportions are Wilson 95% CIs. Detailed definitions, denominators, contingency counts, subgroup results, and descriptive analyses are provided in Supplementary Tables S5–S11. CI, confidence interval; CUQ, Chatbot Usability Questionnaire; FN, false negative; FP, false positive; IQR, interquartile range; SD, standard deviation; TN, true negative; TP, true positive.

Discussion

In this prospective outpatient performance-evaluation study, SpineGPT-2 produced surgeon-facing summaries that exceeded the pre-specified adequacy threshold and supported clinical decision-making in most model-arm encounters. The mean adequacy score was 4.23, significantly above the “safely actionable” threshold of 4.0, and 76.7% of summaries achieved a mean score of at least 4. Diagnostic performance followed a clinically useful pattern: exact top-ranked agreement with the expert-panel consensus was 72.5%, while the panel’s leading diagnosis appeared within the model’s top-three

differential diagnoses in 88.3% of cases. Red-flag detection was also strong when benchmarked against transcript-based expert judgement, with sensitivity of 93.8% and specificity of 96.2%. In addition, the system reduced physician-led history-taking time compared with standard consultation, while maintaining acceptable patient usability. These findings suggest that an agentic AI system can support the earliest stage of spine consultation by collecting a structured pre-consultation history and translating it into a clinically usable summary for the surgeon.

The diagnostic concordance findings suggest that SpineGPT-2 was more successful at identifying the correct diagnostic territory than at always assigning the correct diagnosis the highest priority. This distinction is clinically relevant because the purpose of an initial spine history is rarely to establish a definitive diagnosis in isolation; rather, it is to place the patient into the correct diagnostic pathway and ensure that the surgeon considers the most relevant competing possibilities. The higher top-three concordance therefore supports the system's role as a differential-diagnosis support tool, whereas the lower top-one performance reinforces that diagnostic prioritisation must remain under clinician control. Similarly, the red-flag results indicate that structured AI-assisted history-taking can support early recognition of safety-critical presentations, but cannot replace clinical examination. The lower sensitivity observed against post-examination surgeon verification is consistent with the fact that some neurological or safety-relevant findings may only become apparent during physical examination. Agentic history-taking systems should therefore be understood as pre-consultation safety screens that help bring concerning features to the clinician's attention, rather than as autonomous rule-out tools for serious pathology.

The reduction in physician-led history-taking time represents a practical workflow benefit. The system did not eliminate the communicative work of history-taking; instead, it shifted part of that work to the pre-consultation period and converted patient-reported information into a structured summary before the surgeon entered the encounter. In high-volume spine clinics, even a reduction of a few minutes per patient may translate into meaningful gains in clinic flow, clinician attention, and time available for examination, explanation, and shared decision-making. Patient acceptability further supports feasibility, with a mean CUQ score of 74.6, indicating that most patients were able to engage with the patient-facing interview at an acceptable usability level.

At the same time, the content-audit findings make clear that clinical deployment cannot rely on fluent model output alone. Hallucination was identified in 8.3% of model-arm cases, unsupported clinical inference in 5.0%, and retrieval-provenance label error in 6.7%. Although no hallucination involved a safety-critical red-flag item, the presence of unsupported content supports the need for transparent provenance labels, deterministic safety checks, and mandatory clinician review. Operational safety events were uncommon but not absent, including technical failure in 4.2% of encounters, early exit in 1.7%, and patient-facing hard-rule violation in 1.7%. The relatively low frequency and limited clinical severity of these events support the value of embedding the LLM within an agentic architecture that constrains, labels, and audits its output. However, the fact that such errors were not eliminated altogether indicates that this architecture should not remove the clinician from the loop; rather, it should

make model outputs more transparent, traceable, and easier for the surgeon to verify before they influence clinical reasoning.

The information-gap analysis provides a practical roadmap for system refinement. At least one clinically relevant gap was identified in 27.5% of model-arm encounters, most commonly involving prior treatment and response, functional impact and prior diagnostic workup. These gaps show that the main remaining weakness was incomplete collection of longitudinal and management-relevant history. In spine surgery, decisions often depend on what has already been tried, how symptoms have evolved, which investigations have already been performed, and how the complaint affects function. The next iteration of the system should therefore strengthen these domains, particularly by improving follow-up logic when patients provide vague, incomplete, or non-linear answers.

These findings are clinically relevant because spine consultation depends heavily on converting an undifferentiated complaint into an actionable diagnostic pathway. The system's value was not to replace the surgeon's judgement, but to front-load structured history collection before the surgeon entered the encounter. By organising symptom chronology, pain distribution, neurological complaints, functional limitation, prior treatment response, and systemic features into a concise summary, SpineGPT-2 allowed the consultation to begin from a more organised clinical starting point rather than from unstructured information gathering.

In medical history-taking, agentic systems are increasingly used as workflow components that elicit information, maintain state, apply task constraints, and prepare outputs for downstream clinical review. In the context of medical history-taking, this means that the system is not merely conversing with the patient, but working toward completion of a clinical task under predefined operational constraints.

Earlier consultation agents have already shown that placing an LLM inside a modular architecture can improve the quality and safety of patient interaction[25]. For example, one recent agentic history-taking study compared several leading LLMs with physicians using standardized clinical scenarios. In that study, the agent conducted the interview and produced diagnostic and investigation-related outputs, which were then compared with physician performance. The study showed that agentic systems can perform competitively in structured history-taking tasks[28]. Ren et al. proposed a healthcare-agent framework for medical consultation that combined a dialogue component for safe and effective patient interaction, a memory component for current and previous consultation information, and a processing component for consultation summarisation and report generation[27]. This study also showed, through ablation analyses, that modular agent design can improve inquiry quality, response quality, and safety compared with general LLM use. However, in both studies, evaluation was mainly based on simulated consultations. The Articulate Medical Intelligence Explorer (AMIE) represents an important step in the clinical translation of conversational diagnostic AI. After initial evaluation in simulated OSCE-style consultations, AMIE was subsequently assessed in a prospective ambulatory clinic feasibility study, in which adult patients completed a text-chat interaction before an urgent care appointment. During this pre-encounter interaction, the system collected the clinical history and generated potential diagnoses for

subsequent discussion with the treating clinician[26,29–31]. The study supported the feasibility and acceptability of patient-facing diagnostic dialogue under human safety supervision, and clinicians reported that the AI-generated output improved preparedness for the encounter[29]. Stronger evidence for workflow-level deployment comes from a multicenter randomized trial of a patient-facing preassessment chatbot for primary-to-specialist care transitions[32]. In that study, the system collected history, generated preliminary diagnostic and test-ordering suggestions, and produced referral reports for specialists before the physician encounter. The intervention reduced physician consultation duration and improved physician-perceived care coordination and patient-reported ease of communication, supporting the clinical feasibility of shifting part of the consultation workload upstream into a structured patient-facing interaction[32].

Taken together, the existing literature supports the feasibility of agentic or preassessment-based LLM systems for eliciting clinical information, but also shows that clinical value depends on the surrounding architecture. As conversational diagnostic AI moves closer to clinical care, models are increasingly surrounded by structures that shape its behaviour toward the intended clinical task and make its role accountable within the care pathway[33–35]. SpineGPT-2 follows this broader trajectory but differs in emphasis: selected workflow- and safety-critical functions were implemented in a specialty-specific deterministic orchestration layer around separate patient-facing and surgeon-facing agents.

The architecture of SpineGPT-2 was guided by three deliberate design principles. First, it separated probabilistic clinical flexibility from deterministic safety and workflow control. The value of an LLM in medical history-taking lies in its ability to adapt to different patient expressions, incomplete answers, evolving symptom patterns, and the non-linear nature of clinical storytelling. However, the same probabilistic generation that enables this adaptability means that clinically important behaviours cannot be guaranteed by prompting alone. In a safety-critical setting, some rules must not merely be encouraged; they must be enforced.

SpineGPT-2 was therefore built as a hybrid stochastic-deterministic system. The LLM-based agents provided the adaptive parts of the system: patient-specific questioning during the interview and clinical synthesis after the history had been collected. The deterministic orchestration layer maintained session state, monitored the coverage gate, controlled handoff between agents, recorded the session for audit, and preserved privacy throughout the process. Importantly, it also imposed boundaries that had to remain consistent across encounters: asking only one question per turn, avoidance of giving diagnosis or treatment advice in patient interface and active red-flag screening. These constraints used in SpineGPT-2were specialty- and task-specific, and were not intended to make the interview a rigid checklist. Rather, they defined the boundaries within which the model could remain flexible. They illustrate the broader principle that LLM flexibility becomes clinically useful only when its boundaries are deliberately defined according to the clinical environment in which the system operates, the task assigned to it, the nature of patient interaction, and downstream clinician’s needs. This emphasis on external guardrails is consistent with early clinical evaluations of patient-facing generative AI in digital

health. In an exploratory randomized trial of a generative AI mental health intervention, predefined input and output guardrails prevented the system from returning diagnosis or medical advice, and post-trial review found no failures among 2207 generated responses[36]. Thus, the relevant unit of clinical deployment is not the LLM alone, but the explicitly operational frame that turns a probabilistic model into a bounded clinical component by defining its role, constraints, evidentiary standards, and points of human accountability.

Second, we separated the agent that converses with the patient from the agent that reasons over the collected history. Patient-facing dialogue is inherently shaped by the patient's own framing: patients may present a self-diagnosis, overemphasise a feared symptom, omit clinically important negatives, describe symptoms in emotionally loaded terms, or provide information in a non-linear order. In conventional clinical reasoning, such framing can contribute to anchoring, premature closure, or confirmation bias; cognitive factors are a recognised contributor to diagnostic error in medicine[37]. Recent work on LLM-supported health consultations similarly suggests that patient cognitive bias can influence model reasoning, and that separating the conversational phase from the diagnostic reasoning phase may make the system more robust to bias introduced by the patient's framing[38]. In SpineGPT-2, this concern was addressed architecturally. The PFA was optimised for eliciting information, while the SFA did not participate in the live conversation and reasoned over the locked transcript after the interview was complete. Thus, decoupling dialogue from diagnostic synthesis was intended to reduce framing-related bias while preserving the adaptive advantages of patient-facing conversation.

Third, clinically important statements were grounded by retrieval whenever possible and labelled according to provenance. Retrieval-augmented generation was selected to reduce unsupported generation and lower the risk of hallucination by anchoring clinically relevant claims to an external guideline corpus. However, it was used as one layer of epistemic control, not as a guarantee of correctness: relevant guidance may be absent from the library, the retrieval step may miss the most appropriate source, and the model may still misinterpret retrieved information[39]. Provenance labelling therefore made the status of clinically important statements visible to the surgeon. Retrieved guideline-supported claims carried source attribution; interpretive claims not directly supported by retrieval were marked as model-derived and unverified; and unsupported safety- or triage-relevant claims required surgeon verification. This kept the surgeon-facing summary traceable and aligned with the purpose of augmenting surgeon review rather than functioning as an autonomous diagnostic authority, reinforcing the need for the verification signals built into the system.

The most immediate application of SpineGPT-2 is structured pre-visit intake in high-volume spine clinics, where complete, decision-ready histories are difficult to preserve under time pressure. By shifting part of repetitive history collection upstream and presenting the surgeon with an organised summary, the system may support clinic flow while reserving clinician time for verification, examination, explanation, and management.

Future work should extend evaluation from a single surgeon and centre to a multi-site, multi-surgeon study; assess integration with electronic health records; and test whether the patient-facing agent can safely incorporate multimodal data such as imaging, prior reports, and photographs. Prospective external validation across a broader spectrum of spine pathology will be required before routine adoption.

The principal strengths of this work are its evaluation in real outpatient practice rather than in simulation, its anchoring to a pre-specified and clinically meaningful adequacy threshold, a deterministic safety layer whose guarantees do not depend on model behaviour, a broad multi-dimensional outcome set spanning concordance, red-flag detection, content accuracy, safety and acceptability, and a peer-reviewed foundation on which the agentic redesign was built. The choice of an agentic rather than single-model design is itself supported by evidence that agentic decomposition improves robustness, particularly for safety-critical components[40–41].

Several limitations temper interpretation. The study was conducted with a single surgeon in a secondary-care state hospital rather than a tertiary referral centre. Therefore, the spectrum of spinal pathology may not have been as broad or complex as that seen in high-volume referral institutions, and the findings may not generalise to rarer, atypical, or difficult-to-verbalise spinal conditions. Nevertheless, the cohort included the common degenerative and mechanical presentations that account for a large proportion of routine spine outpatient workload, supporting the pragmatic relevance of the system as a pre-consultation history-taking tool in everyday clinical practice. The primary endpoint was evaluated against an absolute adequacy standard rather than through a direct comparison with clinician-generated histories. The parallel control arm was used not for the primary adequacy analysis, but to represent routine standard history-taking and to provide a comparator for physician-led history-taking time. This was a deliberate design choice in a single-operator setting, intended to avoid carry-over and confounding between model-assisted and standard consultations. However, it means that the study does not establish superiority over usual care for the primary endpoint. Because the expert panel evaluated only model-generated summaries, conventional rater blinding was not possible. The patient-facing interaction was text-based, which may remain unfamiliar to some patients and differs from the natural dynamics of face-to-face clinical history-taking, a caveat also raised in studies of conversational diagnostic AI[26,30]. Retrieval grounding was another limitation. Although central to the system's safety strategy, retrieval cannot guarantee completeness or correctness: relevant guidance may be absent from the indexed corpus, the retrieval step may fail to return the most appropriate passage, and retrieved material may still be misinterpreted or overextended by the model[39].

We designed and evaluated an agentic, safety-first, retrieval-grounded history-taking system in a real spine outpatient clinic; which produced surgeon-facing summaries that met a pre-specified standard of clinical adequacy, with high diagnostic concordance, sensitive red-flag detection, reduced physician-led history-taking time, infrequent and mostly minor errors, and acceptable patient experience. These findings move LLM-based spine anamnesis beyond prompt-level feasibility toward clinical utility as a

supervised pre-consultation support layer in a live specialty setting. Multi-site, multi-surgeon, and multimodal validation is the necessary next step before such a system can be adopted in routine care

Methods

Study design

This was a prospective performance-evaluation study conducted in a single spine outpatient clinic. The study assessed whether the surgeon-facing output of an AI-based history-taking system met pre-specified clinical adequacy targets in routine outpatient practice. The study was approved by the Tokat Gaziosmanpaşa University Institutional Ethics Committee (no. 26-MOBAEK-169; date: 12/05/2026), and all participants provided written informed consent before enrolment. The study is reported in accordance with the DECIDE-AI reporting guidelines[42].

The SpineGPT-2 system

SpineGPT-2 was implemented as an agentic history-taking system composed of three sequential components: a patient-facing agent (PFA), a deterministic orchestration layer, and a surgeon-facing agent (SFA). The PFA conducts the clinical interview directly with the patient through a text-based chat interface, gathering a structured spine history and calling clinical instruments when relevant. The orchestration layer is a conventional rule-based program written in Python and contains no language model itself. It supervises the interview, mediates the handoff to the second agent, persists data, and enforces a set of safety-critical behaviours that cannot be left to the unpredictable behaviour of a language model. The resulting architecture separated adaptive patient interviewing from deterministic workflow and safety control and from downstream clinical synthesis, as shown in Fig. 5.

Generation parameters were held constant throughout the study. For the PFA, temperature was set to 0.4 and the maximum output length to 2,048 tokens; for the SFA, temperature was set to 0.2 and the maximum output length to 8,192 tokens. For both agents, top-p was set to 1.0, and frequency and presence penalties were set to 0. All API-based inference used to generate the study dataset was performed between May and June 2026, corresponding to the patient recruitment period. Both agents were accessed through the OpenAI API using the model identifier gpt-4.1.

Patient-facing agent (PFA). The PFA conducts an adaptive clinical interview. Its system prompt frames the task as clinical reasoning over a partial picture, in which each answer is treated as new information and the next question is chosen to most reduce diagnostic uncertainty. To prevent this adaptive behaviour from omitting clinically essential content, the prompt also defines a set of mandatory minimum history elements (patient demographics, the chief complaint and its main properties, such as onset, duration, location, radiation, severity, aggravating and relieving factors, a separately conducted neurological inquiry, functional impact, relevant medical background, and a systemic red-flag screen; (Supplementary Table S1)). Before closing, the agent is instructed to run an explicit self-check against

this list and to ask the patient whether anything further should be added. At the start of each encounter the patient selected either the local language or English as the interface language, which was then locked for the remainder of the session; the interview was conducted in this language, whereas the surgeon-facing report was always generated in English. The interview followed a clinical order in which the chief complaint was fully characterised, including its location, severity, character, radiation, and aggravating and relieving factors, before any additional complaints were explored.

The PFA operates under a set of behavioural rules governed by the orchestration layer. It asks one question at a time, phrases all questions in plain language and follows up on any clinically significant disclosure. It does not provide a diagnosis, does not recommend treatment, and may close the interview only after completing the baseline self-check and receiving the patient's final confirmation. The agent is also methodologically blinded to the adequacy rubric against which its output will later be scored, so that it cannot tailor its interview behaviour to the evaluation instrument. The PFA system prompt and runtime implementation are provided in Supplementary Software.

Orchestration layer. The orchestration layer was deterministic Python code with no language-model component. It governed the parts of the workflow that had to behave predictably across encounters, including session state, patient-visible output control, data handoff, closure logic, privacy protection, retrieval calls, and audit logging. It maintained a continuously updated encounter record, preserving every patient–agent exchange, system event, and retrieved evidence block so that the interview could be reviewed, audited, and summarised from a complete session history. It maintained a continuously updated record of the encounter, preserving every patient-agent exchange and any retrieved information so that the interview could be reviewed, audited, and summarised from a complete session history. On every conversational turn, the orchestrator supplied the PFA with the relevant session context and inserted a short reminder reasserting the behavioural rules that the PFA had to follow. In addition, every PFA response was intercepted by a rule-based gate before being delivered to the patient. This gate provided protection beyond the agent's own instruction set by checking patient-facing output for prohibited behaviours such as diagnostic assertions or treatment recommendations. A baseline level of history completeness was maintained through a deterministic coverage gate without converting the interview into a fixed checklist. The mandatory minimum history elements were derived from our previously published SpineGPT framework[23]. During the encounter, the agent selected each question adaptively based on the patient's preceding answer and the evolving clinical context. When the agent attempted to close the interview, the orchestrator reviewed the transcript and verified whether each mandatory item had been addressed or explicitly marked as not applicable. If the history remained below this baseline threshold, closure was blocked, the outstanding items were returned to the agent, and the interview resumed until the check was satisfied or the predefined reopening limit was reached. Thus, clinical judgement regarding applicability remained with the agent, whereas baseline adequacy auditing and termination control were handled deterministically by the orchestrator. A safeguard capped the number of times closing could be reopened so that the interview always terminated in a controlled way. A second deterministic mechanism operated alongside the coverage gate to promote focused follow-up of clinically important disclosures while preserving adaptive question selection. At each turn, the

orchestrator screened the patient's most recent response for a predefined set of clinically significant signals, including progressive weakness, gait or balance disturbance, night or rest pain, prior malignancy, recent infection, unintentional weight loss, sphincter dysfunction, recent trauma, functional decline, and features suggestive of reduced bone quality. When a previously untriggered signal was detected, the orchestrator issued a targeted, single-use instruction asking the PFA to characterise that finding before continuing the interview. This instruction was non-blocking and did not otherwise determine question order or interview termination; each signal could trigger the instruction only once per encounter. The PFA therefore retained control over subsequent question selection without converting the interview into a fixed script.

Several further functions are handled within the orchestration layer. Each patient is assigned an anonymous study token in the format SPG-XXXX, and the system does not collect or store personal identifying information. After interview closure, the orchestrator locked the session, prepared the structured encounter package for the SFA, initiated retrieval from the fixed guideline corpus, and stored the generated report with the session record. Surgeon access to the generated clinical summary was restricted through a separate PIN-protected interface. The same deterministic layer also audited the surgeon-facing report before release, verifying that clinically meaningful statements carried the provenance label appropriate to their content; the operation of this validator is described under SFA. The implementations of the deterministic router, session-management module, coverage gate, deepen-on-signal mechanism, and operational metrics logger are provided in Supplementary Software.

Surgeon-facing agent (SFA). After the interview was completed, the orchestrator locked the session and transferred a structured encounter package to the SFA. This package included the verbatim interview transcript, interview duration and the retrieved guideline material described below. To preserve independent clinical synthesis, the SFA receives the original transcript rather than a summary produced by the PFA and generates its report only from the locked encounter package and the retrieved guideline material described below.

The SFA's output follows a three-section structure: Patient's Story, a narrative reconstruction of the history; Diagnostic Impression, containing a step-by-step reasoning chain and a ranked differential diagnosis with supporting and opposing evidence for each candidate; and Recommended Next Steps, including specific physical-examination maneuvers, suggested laboratory or imaging investigations, and triage. Red flag findings are reported as identified or not explored. This prevents an item that was not reached during the interview from being misread as an item that was actively ruled out.

Before report generation, the system queried a fixed English-language clinical reference corpus of 31 indexed records: 17 full-text AO Spine or AO Spine-affiliated publications, three additional spine-related publications indexed from abstracts only, and 11 full-text North American Spine Society guideline documents (listed in Supplementary Information). The indexed corpus is searched for safety-relevant findings, differential diagnoses, and diagnostic workup, with workup queries anchored to the leading diagnostic candidates. Retrieved passages are relevance-filtered, boilerplate-screened, and inserted into

the agent's context as labelled evidence blocks. Retrieval used dense sentence-level embeddings, with indexed passages ranked by cosine similarity. Passages below a prespecified minimum relevance threshold were excluded, and the corpus and its index remained unchanged throughout the study.

The SFA report used a provenance policy in which source labelling depended on the clinical importance of each statement. Statements directly supported by retrieved guideline material were marked with an asterisk and a numbered reference linked to the report's reference list. Interpretive or contextual statements without retrieved support were labelled as model-derived and unverified, indicating that they relied on the model's general training knowledge and required clinician judgement. Statements with potential implications for patient safety or triage were not allowed to rely silently on model knowledge. They had to be either supported by a retrieved source or explicitly marked as requiring surgeon verification. Immediately after report generation, a deterministic validator in the orchestration layer checked whether these rules had been applied correctly and whether every citation corresponded to a source actually retrieved for that session. All corrections were stored as session events, allowing provenance-policy violations to be quantified across the study as an objective measure of system adherence. The SFA system prompt, report-generation implementation, and deterministic provenance-validation logic are provided in Supplementary Software. Figure 6 presents a simulated encounter created solely to illustrate the interview-to-summary workflow; the complete transcript and corresponding report are provided in Supplementary Information.

Participants

We enrolled 240 patients attending the spine outpatient clinic during the study period, with 120 patients assigned to the model arm and 120 to the control arm. Recruitment was conducted during the prespecified study period. Allocation was performed at the clinic-day level using an alternating day-block design: entire clinic days were designated in advance as either model days or control days. Eligible patients were adults presenting with a spine-related complaint, who were able to use a tablet-based chat interface without assistance and provided informed consent. Patients were excluded if they required immediate emergency management or were unable to interact with the interface independently because of severe visual, hearing, cognitive, or other functional limitations. On model days, patients completed the SpineGPT-2 interview in the waiting area before the consultation; the surgeon-facing summary was then generated and made available to the treating surgeon, who subsequently examined and managed the patient according to usual clinical practice. On control days, the same surgeon obtained the history, examined the patient, and managed the consultation in the standard manner without access to a SpineGPT-2 summary. Patient recruitment and all API-based inference used to generate the patient-facing interviews and surgeon-facing reports analysed in this study were conducted concurrently during May and June 2026. The arm-specific workflows and physician-led history-taking intervals are illustrated in Fig. 7.

Outcomes

Primary outcome: Adequacy of the surgeon-facing summary. The primary outcome was the clinical adequacy of the surgeon-facing summary in the model arm. Each summary was evaluated by an independent expert panel of specialists from neurosurgery or orthopaedic surgery backgrounds. For each case, panellists rated five statements on a five-point Likert scale, from 1 = strongly disagree to 5 = strongly agree. Each item was phrased to assess whether, after reading the summary, the rater could perform the relevant clinical task with confidence:

Q1. I can formulate an appropriate differential diagnosis.

Q2. I can determine the next diagnostic steps.

Q3. I can determine the appropriate triage category for this patient.

Q4. I can identify whether safety-critical features are present or absent.

Q5. The history is sufficiently informative to proceed with the consultation without needing to repeat the full history.

Each summary was rated independently by three panellists. For each item, the mean of the three ratings was used in the primary analysis, and the five resulting item scores were averaged to obtain the per-case overall adequacy score. We pre-specified a threshold of 4.0 as the 'safely actionable' bar: because 4 corresponds to 'agree, with confidence,' a mean above 4 indicates that, on average, the summary was judged sufficient to act on confidently.

Diagnostic concordance and accuracy. For each case, the expert panel also recorded its ranked differential diagnoses (Q6) and recommended next diagnostic or management steps (Q7) as free text. To prevent the panel's judgement from being influenced by model-generated diagnostic suggestions, panellists recorded these responses before viewing the SFA's Diagnostic Impression and Recommended Next Steps sections. Concordance between the model's ranked differential diagnosis and the panel consensus was assessed descriptively. Exact agreement with the panel's top-ranked diagnosis was used as the primary diagnostic concordance measure, while inclusion of the panel's leading diagnosis within the model's top three differentials was reported as a secondary measure of differential coverage. We also report inter-rater agreement among the panellists, providing a reference point for the level of diagnostic agreement that could reasonably be expected in this clinical setting. A diagnostic-accuracy sub-analysis was performed only in the subset of patients for whom the final diagnosis was confirmed by imaging, laboratory testing, or another clinically accepted diagnostic investigation. In these reference-standard cases, we report how often the confirmed diagnosis was ranked first by the model and how often it appeared anywhere within the model's top three differential diagnoses.

Red-flag detection. The primary analysis compared the SFA red-flag report with the expert panel's red-flag judgement based on the interview transcript. Sensitivity and specificity were calculated against this reference standard and reported with 95% Wilson confidence intervals. Given the expected low prevalence of true red flags, these estimates were interpreted descriptively. As an exploratory secondary

analysis, the SFA red-flag report was also compared with the treating surgeon's post-examination verification, allowing assessment of findings that may not be fully derivable from history alone, such as examination-only neurological deficits. The a priori red-flag categories included suspected acute or progressive neurological deficit, previous malignancy, systemic infection markers, night pain with systemic features, new-onset back pain at advanced age, recent significant trauma, and immunosuppression (Supplementary Table S2).

Consultation-time efficiency. As a pre-specified secondary outcome, and following the same approach as our prior work, we compared the duration of the physician's history-taking between the two arms. This outcome was intended to assess whether a surgeon-facing summary available at the start of the encounter reduced the time the surgeon needed to spend establishing the history. Consultation time was measured by the clinic secretary using a stopwatch. In the control arm, timing started when the surgeon began taking the history and ended when the history was complete. In the model arm, patients first completed the SpineGPT-2 interview in the waiting area, and the generated summary was available before the consultation. Timing started when the surgeon began reviewing the summary and asking any additional history questions, and ended when the history was complete. Thus, in both arms, the measured interval represented only the physician-led portion of history-taking. The time patients spent completing the pre-visit SpineGPT-2 interview was recorded separately from the system log and was not included in this outcome, because it occurred before the physician encounter and did not consume surgeon time.

Content accuracy and hallucination. In an exploratory quality-control analysis, a senior author reviewed model-arm cases by comparing the system-log interview transcript with the surgeon-facing summary. The review focused on whether the summary contained hallucinated content, defined as clinically relevant information stated in the summary but not supported by the interview transcript; unsupported clinical inference, defined as a diagnostic or management-related conclusion not justified by the recorded history or retrieved evidence; and citation errors, defined as numbered guideline references that did not correspond to the retrieved sources. Results were reported descriptively as the number and proportion of reviewed cases containing each issue.

Operational safety events. As all model-arm patients were subsequently seen by the treating surgeon and no clinical decision relied solely on system output, monitoring focused on operational and interaction-safety events rather than clinical adverse events. We tracked three predefined categories: technical failures preventing normal interview completion; patient-initiated early exits before structured closure; and patient-facing hard-rule violations, defined as any diagnostic assertion or treatment recommendation that reached the patient in the chat interface. Hard-rule violations were identified through post-hoc review of all model-arm chat transcripts, and the total number of events was reported for each category.

Exploratory information-gap analysis. When the expert panel judged that clinically relevant information was missing, the missing item was assigned to one of eight predefined spine-history domains: pain

characterisation, pain trajectory, systemic safety screening, neurological inquiry, functional impact, prior treatment, prior diagnostic workup, and relevant medical history. This analysis was used descriptively to identify which parts of the spine history were most often under-collected.

Patient experience and acceptability. Patient experience was assessed only in the model arm as evidence of the acceptability of the agent interaction, using the Chatbot Usability Questionnaire (CUQ), a validated 16-item instrument designed specifically for conversational agents. No patient-experience questionnaire was administered in the control arm; the CUQ results are therefore reported descriptively as acceptability evidence for the patient-facing system.

Statistical Analysis

The sample size was based on the primary adequacy outcome in the model arm. We planned to test whether the mean summary adequacy score exceeded the prespecified threshold of 4.0 using a one-sample, one-sided t-test. Assuming an expected mean adequacy score of 4.3, a standard deviation of 0.8, a one-sided α of 0.05, and 80% power, the minimum required sample size for the primary one-sample t-test was 46 model-arm cases. We increased the sample to 120 patients per arm to improve the precision of descriptive estimates across secondary and exploratory outcomes, including diagnostic concordance, red-flag detection, operational safety events, and content accuracy, rather than to formally power each secondary endpoint.

The primary analysis tested whether the mean adequacy score exceeded 4.0 using a one-sample, one-sided t-test. We reported the mean adequacy score, 95% confidence interval, t statistic, one-sided P value, and Cohen's d. Inter-rater reliability for the overall adequacy score across the three panellists was assessed using a two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficient (ICC[2,3]). Secondary and exploratory outcomes were summarised descriptively. Continuous variables were reported as means with standard deviations or medians with interquartile ranges, as appropriate. The between-arm difference in physician-led history-taking time was assessed with the Mann–Whitney U test, reported with medians and interquartile ranges and a rank-biserial correlation as the effect size. Proportions, including diagnostic concordance, red-flag sensitivity and specificity, operational-safety-event rates, and content-accuracy findings, were reported with 95% Wilson confidence intervals. Inter-rater reliability for the adequacy rubric was assessed using the intraclass correlation coefficient. Inter-rater agreement for the categorical primary-diagnosis assignments made independently by the three panellists was assessed using Fleiss' κ. Items judged not applicable were excluded from the relevant item-level denominator. Analyses were performed in Python using numpy, pandas, scipy, and statsmodels, with a two-sided α of 0.05 used for any secondary hypothesis tests.

Declarations

Competing interests:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

B.A.K. conceived the study, developed the study methodology and wrote the software, coordinated the project, and drafted the manuscript. A.K. conducted the clinical study, managed the outpatient workflow, performed the clinical assessment and usual-care history taking, and contributed to data acquisition and interpretation. T.E., Ö.Ç., and İ.S. independently evaluated the model-generated outputs and contributed to clinical validation and interpretation of the findings. A.A. contributed to the statistical methodology, performed the formal analyses, and assisted with interpretation and visualisation of the results. SŞ provided senior clinical supervision, contributed to study methodology, reviewed the model outputs for clinical content and safety, and assisted with interpretation of the findings. All authors critically revised the manuscript, approved the final version, and accept responsibility for the integrity of the work.

Acknowledgements:

None

Data Availability

Data availability: The participant-level transcripts, generated surgeon-facing reports, operational logs, and token-linked encounter records are not publicly available because of participant privacy, consent limitations, and institutional ethics restrictions. Aggregate results supporting the findings are reported in the Article and Supplementary Information. The simulated transcript and corresponding report as well as the document-level retrieval-corpus manifest are provided as Supplementary Information.Code availability: A blinded core-code review package is provided as Supplementary Software. The package contains the patient-facing and surgeon-facing agent implementations and system prompts, deterministic orchestration modules, retrieval and provenance-validation code, and dependency specifications. The code is additionally maintained in a private repository (SpineGPT-2), which will be made publicly available and archived with a persistent DOI upon acceptance. To protect participant confidentiality and comply with third-party copyright restrictions, the package does not include patient-level data, real session directories, interview transcripts, complete surgeon reports, operational log files and copyrighted guideline full text, embeddings, or the persistent vector index. The complete unblinded archival implementation is retained by the authors and may be considered for controlled access subject to institutional, ethical, intellectual-property, and third-party licensing restrictions.

Code availability:

A blinded core-code review package is provided as Supplementary Software. The package contains the patient-facing and surgeon-facing agent implementations and system prompts, deterministic orchestration modules, retrieval and provenance-validation code, and dependency specifications. The code is additionally maintained in a private repository (SpineGPT-2), which will be made publicly available and archived with a persistent DOI upon acceptance. To protect participant confidentiality and comply with third-party copyright restrictions, the package does not include patient-level data, real session directories, interview transcripts, complete surgeon reports, operational log files and copyrighted guideline full text, embeddings, or the persistent vector index. The complete unblinded archival implementation is retained by the authors and may be considered for controlled access subject to institutional, ethical, intellectual-property, and third-party licensing restrictions.

References

1. Peterson, M. C., Holbrook, J. H., Von Hales, D., Smith, N. L. & Staker, L. V. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J. Med. 156, 163–165 (1992).
2. Sandler, G. The importance of the history in the medical clinic and the cost of unnecessary tests. Am. Heart J. 100, 928–931 (1980).
3. Suri, P., Rainville, J., Kalichman, L. & Katz, J. N. Does this older adult with lower extremity pain have the clinical syndrome of lumbar spinal stenosis? JAMA 304, 2628–2636 (2010).
4. Cook, C. J. et al. Systematic review of diagnostic accuracy of patient history, clinical findings, and physical tests in the diagnosis of lumbar spinal stenosis. Eur. Spine J. 29, 93–112 (2020).
5. Shultz, S., Averell, K., Eickelman, A., Sanker, H. & Donaldson, M. B. Diagnostic accuracy of self-report and subjective history in the diagnosis of low back-related leg pain. Spine J. 15, 979–993 (2015).
6. Verhagen, A. P., Downie, A., Popal, N., Maher, C. & Koes, B. W. Most red flags for malignancy in low back pain guidelines lack empirical support: a systematic review. Pain 158, 1860–1868 (2017).
7. Henschke, N. et al. Red flags to screen for malignancy in patients with low-back pain. Cochrane Database Syst. Rev. CD008686 (2013).
8. Downie, A. et al. Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. BMJ 347, f7095 (2013).
9. Dionne, N. et al. What is the diagnostic accuracy of red flags related to cauda equina syndrome? A systematic review. Musculoskelet. Sci. Pract. 42, 125–133 (2019).
10. Oliveira, C. B. et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. Eur. Spine J. 27, 2791–2803 (2018).
11. Verhagen, A. P., Downie, A., Popal, N., Maher, C. & Koes, B. W. Red flags presented in current low back pain guidelines: a review. Eur. Spine J. 25, 2788–2802 (2016).
12. Bath, B., Grona, S. L. & Janzen, B. A spinal triage programme delivered by physiotherapists in collaboration with orthopaedic surgeons. Physiother. Can. 64, 356–366 (2012).

13. Hourigan, P. G. & Weatherley, C. R. Initial assessment and follow-up by a physiotherapist of patients with back pain referred to a spinal clinic. J. R. Soc. Med. 87, 213–214 (1994).
14. Weatherley, C. R. & Hourigan, P. G. Triage of back pain by physiotherapists in orthopaedic clinics. J. R. Soc. Med. 91, 377–379 (1998).
15. Robarts, S., Stratford, P., Kennedy, D., Malcolm, B. & Finkelstein, J. Evaluation of an advanced-practice physiotherapist in triaging patients with lumbar spine pain: surgeon–physiotherapist level of agreement and patient satisfaction. Can. J. Surg. 60, 266–272 (2017).
16. Wong, W. S. Y., Sun, C. M., Koh, H. Y., Tan, L. R. H. & Huang, Y. Reducing wait times and medical costs for patients: the physiotherapy-led Spine Triage and Rehabilitation clinic. BMJ Open Qual. 13, e002670 (2024).
17. Singhal, K. et al. Toward expert-level medical question answering with large language models. Nat. Med. 31, 943–950 (2025).
18. Gilson, A. et al. How does ChatGPT perform on the USMLE? Implications for medical education and knowledge assessment. JMIR Med. Educ. 9, e45312 (2023).
19. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 333, 319–328 (2025).
20. Artsi, Y. et al. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front. Digit. Health 7, 1659134 (2025).
21. Leypold, T., Schäfer, B., Boos, A. M. & Beier, J. P. Artificial intelligence-powered hand surgery consultation: GPT-4 as an assistant in a hand surgery outpatient clinic. J. Hand Surg. Am. 49, 1078–1088 (2024).
22. Leypold, T., Bahm, J., Beier, J. P. et al. Evaluating ChatGPT o1's capabilities in peripheral nerve surgery: advancing artificial intelligence in clinical practice. World Neurosurg. 196, 123753 (2025).
23. Kaplan, A., Aslantaş Kaplan, B., Aydilek, A. et al. Spine patient history-taking: a two phase study of a prompt-engineered large language model assistant. Eur. Spine J. https://doi.org/10.1007/s00586-026-10094-0 (2026).
24. Hinostroza Fuentes, V. G., Karim, H. A., Tan, M. J. T. & AlDahoul, N. AI with agency: a vision for adaptive, efficient, and ethical healthcare. Front. Digit. Health 7, 1600216 (2025).
25. Collaco, B. G., Haider, S. A., Prabha, S. et al. The role of agentic artificial intelligence in healthcare: a scoping review. npj Digit. Med. 9, 345 (2026).
26. Tu, T., Schaekermann, M., Palepu, A. et al. Towards conversational diagnostic artificial intelligence. Nature 642, 442–450 (2025).
27. Ren, Z. et al. Healthcare agent: eliciting the power of large language models for medical consultation. npj Artif. Intell. 1, 24 (2025).
28. Steinbuch, S. et al. Comparative performance of agentic AI and physicians in taking clinical history across leading large language models. Preprint at medRxiv https://doi.org/10.64898/2026.01.23.26344723 (2026).

29. Brodeur, P., Koshy, J. M., Palepu, A. et al. A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. Preprint at arXiv https://doi.org/10.48550/arXiv.2603.08448 (2026).
30. Saab, K., Palepu, A., Schaekermann, M. et al. Advancing conversational diagnostic AI with multimodal reasoning. Nat. Med. https://doi.org/10.1038/s41591-026-04371-0 (2026).
31. Liévin, V., Palepu, A., Weng, W.-H. et al. Towards conversational AI for disease management. Nature https://doi.org/10.1038/s41586-026-10764-5 (2026).
32. Tao, X., Zhou, S., Ding, K. et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nat. Med. 32, 934–942 (2026).
33. Haider, S. A. et al. The development and evaluation of a retrieval-augmented generation large language model virtual assistant for postoperative instructions. Bioengineering 12, 1219 (2025).
34. Umer, F., Shaikh, M. M. & Ur Rahman, A. Evaluating the safety of large language models in healthcare and dentistry: adversarial testing approaches. BDJ Open 12, 70 (2026).
35. Agweyu, A., Mwaniki, P., Musau, W. et al. Safety of a large language model-based clinical decision support system in African primary healthcare. Nat. Health 1, 607–618 (2026).
36. Campellone, T. R., Flom, M., Montgomery, R. M. et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. J. Med. Internet Res. 27, e67365 (2025).
37. Saposnik, G., Redelmeier, D., Ruff, C. C. & Tobler, P. N. Cognitive biases associated with medical decisions: a systematic review. BMC Med. Inform. Decis. Mak. 16, 138 (2016).
38. Zuo, Y., Wan, Q. & Wang, S. Patient cognitive bias in large language model-supported health consultations: simulation-based comparative study. J. Med. Internet Res. 28, e85770 (2026).
39. Zhang, G., Xu, Z., Jin, Q. et al. Leveraging long context in retrieval augmented language models for medical question answering. npj Digit. Med. 8, 239 (2025).
40. AlSammarraie, A., Al-Saifi, A., Kamhia, H., Aboagla, M. & Househ, M. Development and evaluation of an agentic LLM-based RAG framework for evidence-based patient education. BMJ Health Care Inform. 32, e101570 (2025).
41. Lee, D. T., Yen, S., Smits, G. et al. Use of AI agents to assess preoperative frailty in cancer patients. npj Digit. Surg. 1, 9 (2026).
42. Vasey, B., Nagendran, M., Campbell, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat. Med. 28, 924–933 (2022).

Figures

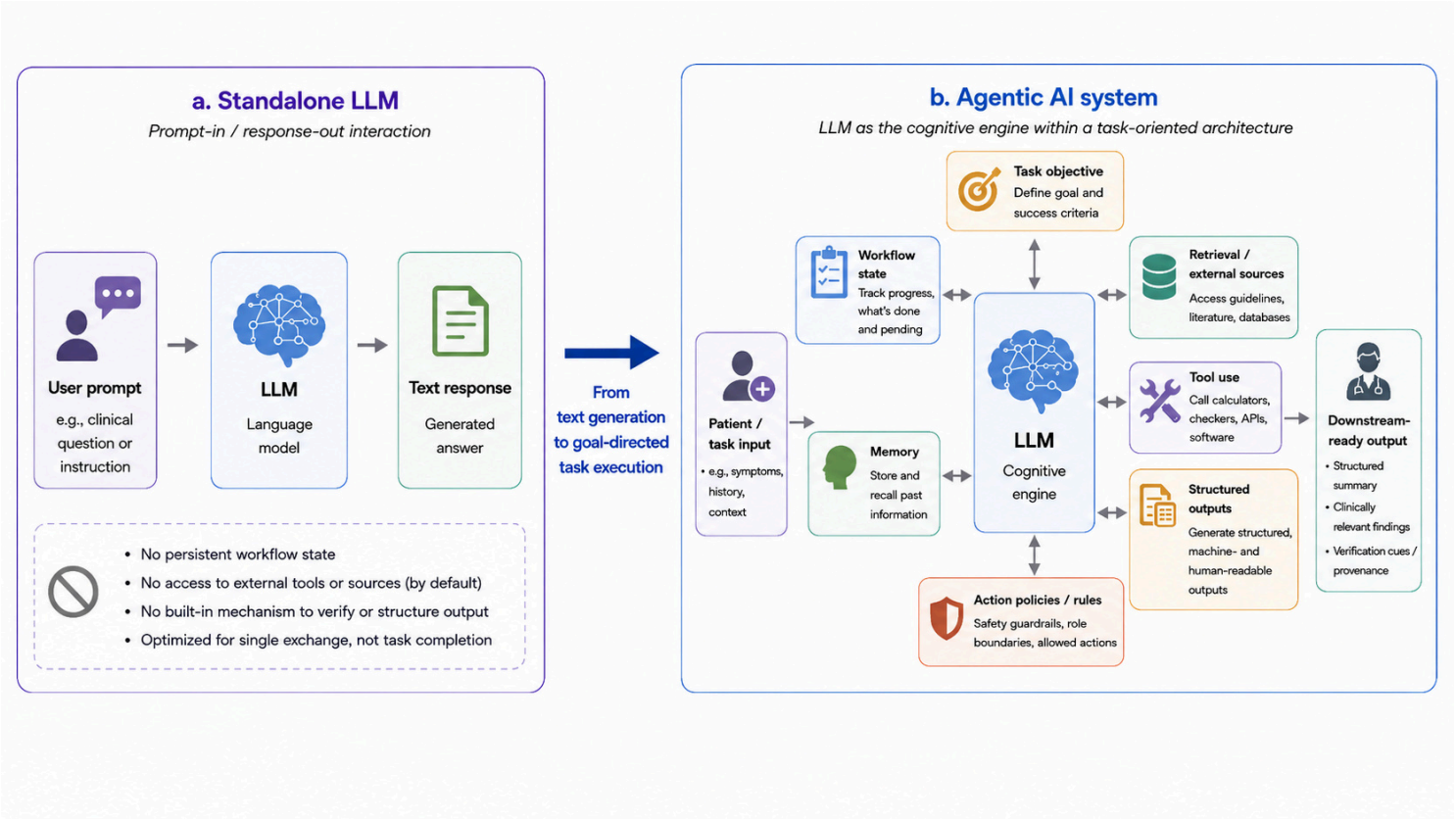


Figure 1

a. A standalone large language model typically receives a prompt and returns a text response without persistent task state, integrated external resources, or workflow-level output controls. b. An agentic AI system embeds the large language model as a cognitive engine within a task-oriented architecture that can maintain workflow state, access memory and retrieved evidence, call external tools, apply predefined action policies, and generate structured outputs for downstream use. This is a conceptual schematic rather than an exact representation of the SpineGPT-2 architecture, which is shown in Figure 5. API, application programming interface; LLM, large language model.

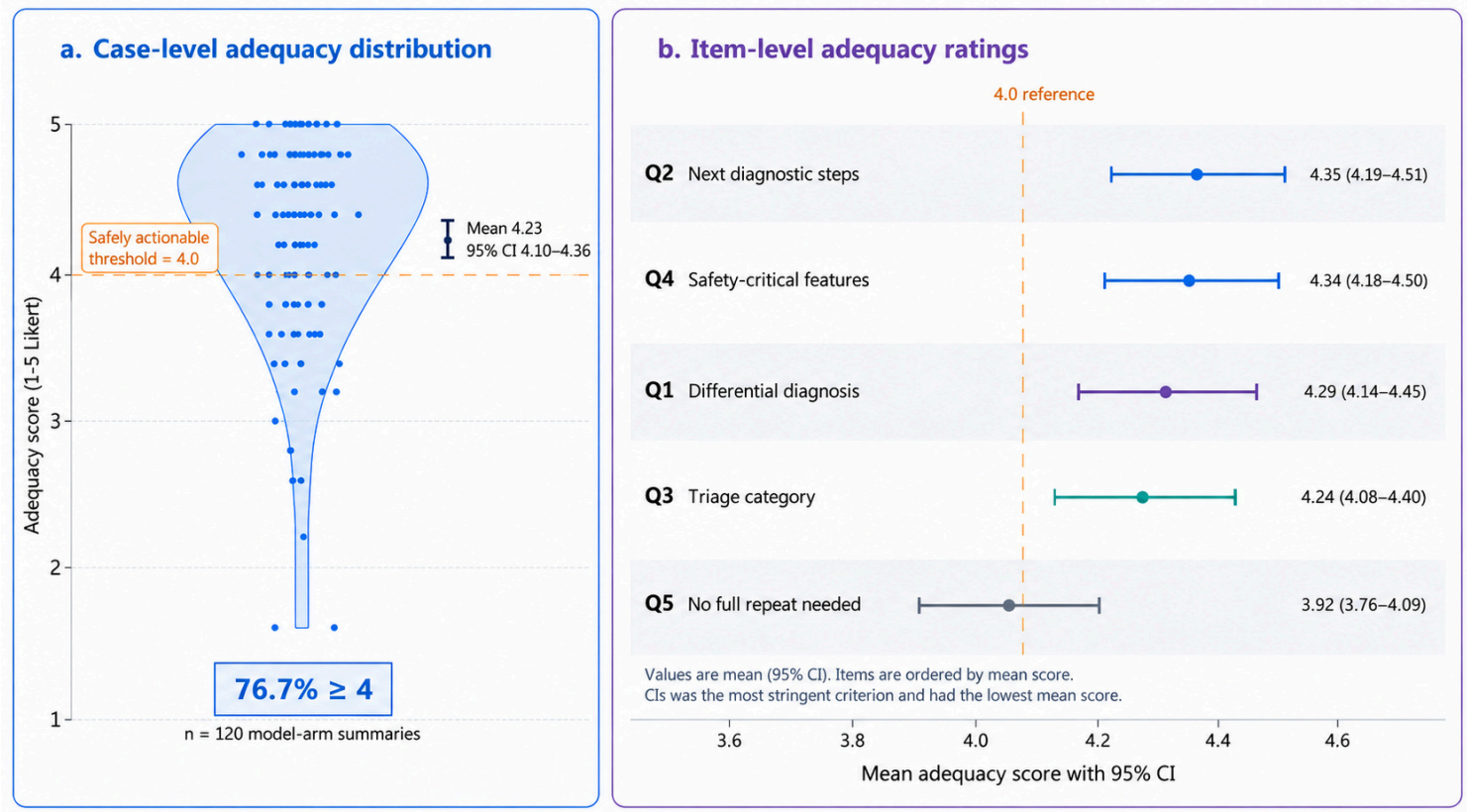


Figure 2

a. Distribution of the per-case mean adequacy scores across 120 model-arm summaries. Each point represents one case, and the violin indicates the distribution of scores. The dashed horizontal line denotes the prespecified safely actionable threshold of 4.0. The overall mean was 4.23 (95% confidence interval, 4.10–4.36), and 76.7% of summaries achieved a mean adequacy score of at least 4. b. Mean item-level adequacy ratings with 95% confidence intervals, ordered by mean score. Q5, which assessed whether the consultation could proceed without repeating the full history, was the most stringent item and had the lowest mean score. Ratings were made on a five-point Likert scale from 1, strongly disagree, to 5, strongly agree. CI, confidence interval.

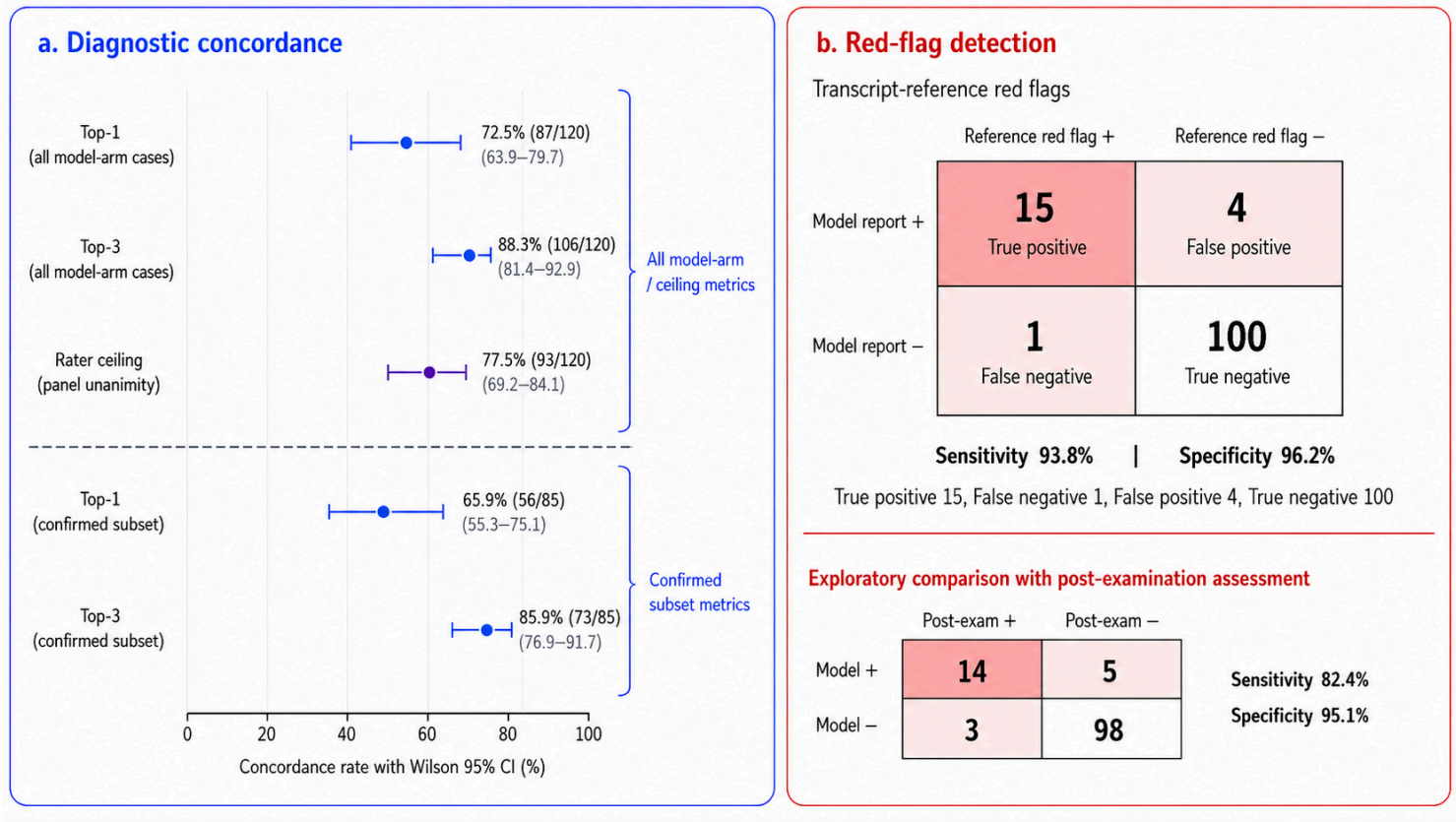


Figure 3

a. Top-1 and top-3 diagnostic concordance with expert-panel consensus across all 120 model-arm cases, panel unanimity as a contextual benchmark, and diagnostic concordance within the 85-case subset with an objective reference diagnosis. Points show proportions, and horizontal bars show Wilson 95% confidence intervals. In the confirmed-diagnosis subset, the confirmed diagnosis was ranked first in 56 of 85 cases (65.9%) and appeared within the model's top-three differential diagnoses in 73 of 85 cases (85.9%). b. Confusion matrices for red-flag detection against the expert panel's transcript-based judgement in the primary analysis and against the treating surgeon's post-examination verification in the exploratory analysis. Sensitivity and specificity are shown for each reference standard. CI, confidence interval; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

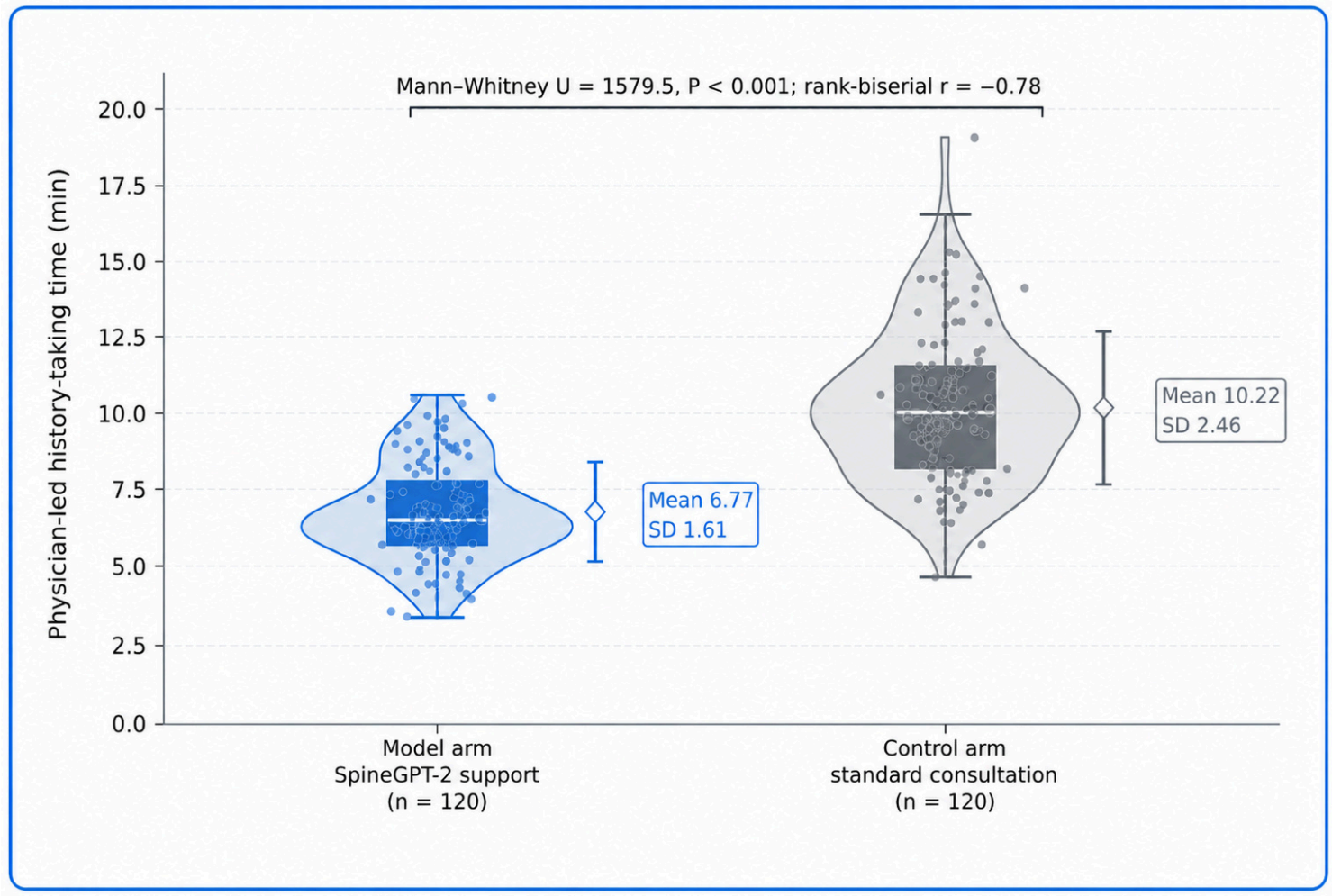


Figure 4

Violin plots show the distribution of physician-led history-taking times for 120 patients in each arm. Individual points represent patients, boxes indicate the median and interquartile range, and diamonds with error bars indicate the mean and standard deviation. In the model arm, the measured interval included surgeon review of the SpineGPT-2 summary and any additional history questions but excluded the pre-visit patient-facing interview. Groups were compared using a two-sided Mann–Whitney U test (U = 1579.5, P < 0.001; rank-biserial r = −0.78). IQR, interquartile range; SD, standard deviation.

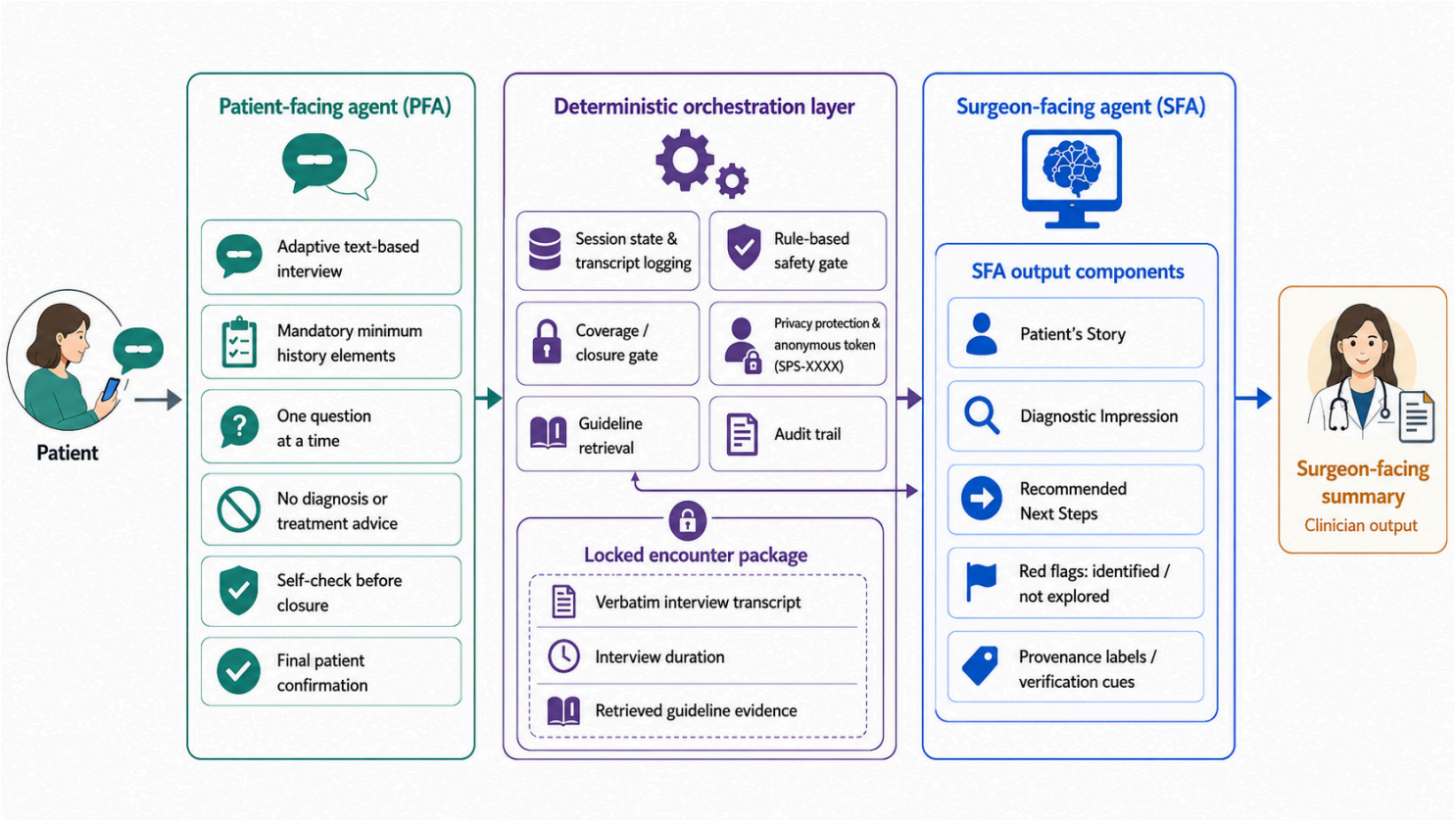


Figure 5

The patient-facing agent conducts an adaptive text-based interview under predefined behavioural constraints and mandatory minimum history requirements. The deterministic orchestration layer maintains session state, records the transcript, enforces patient-facing safety and closure rules, protects privacy through an anonymous study token, performs guideline retrieval, and preserves an audit trail. After interview closure, the encounter is locked and a structured package containing the verbatim transcript, interview duration and retrieved guideline evidence is transferred to the surgeon-facing agent. The surgeon-facing agent generates the Patient's Story, Diagnostic Impression, and Recommended Next Steps and includes red-flag status and provenance or verification cues for clinician review. PFA, patient-facing agent; SFA, surgeon-facing agent.



Figure 6

a. Patient-facing interface lifecycle, including language selection, anonymous study-token generation, interview initiation and completion, and PIN-protected access to the generated report by the surgeon. The study token is masked as SPG-XXXX for display. b. Selected excerpts from the patient-facing interview illustrating adaptive questioning. Questions are shown in chronological order, and ellipses indicate intervening exchanges omitted for readability. The “Reason/Next Focus” column shows how

each patient response changed the subsequent information-gathering objective, rather than the interview proceeding as a fixed checklist. The highlighted exchange following the patient's question, “Do you think it might be cancer?”, illustrates the diagnostic-boundary safeguard: the system acknowledges the concern and continues clinically relevant information gathering without providing a diagnosis. Deterministic post-interview checkpoints included coverage assessment, extraction of clinically significant findings, red-flag screening, transcript locking, guideline retrieval, and provenance validation. c. Abridged surgeon-facing agent output generated from the locked encounter record, comprising the Patient's Story, stepwise clinical reasoning, ranked differential diagnosis, red-flag assessment, and recommended next steps. Evidence labels and numbered source boxes identify statements supported by passages retrieved from two North American Spine Society guideline documents. Dialogue and report excerpts were shortened for display without changing their clinical meaning. PFA, patient-facing agent; SFA, surgeon-facing agent; NASS, North American Spine Society.

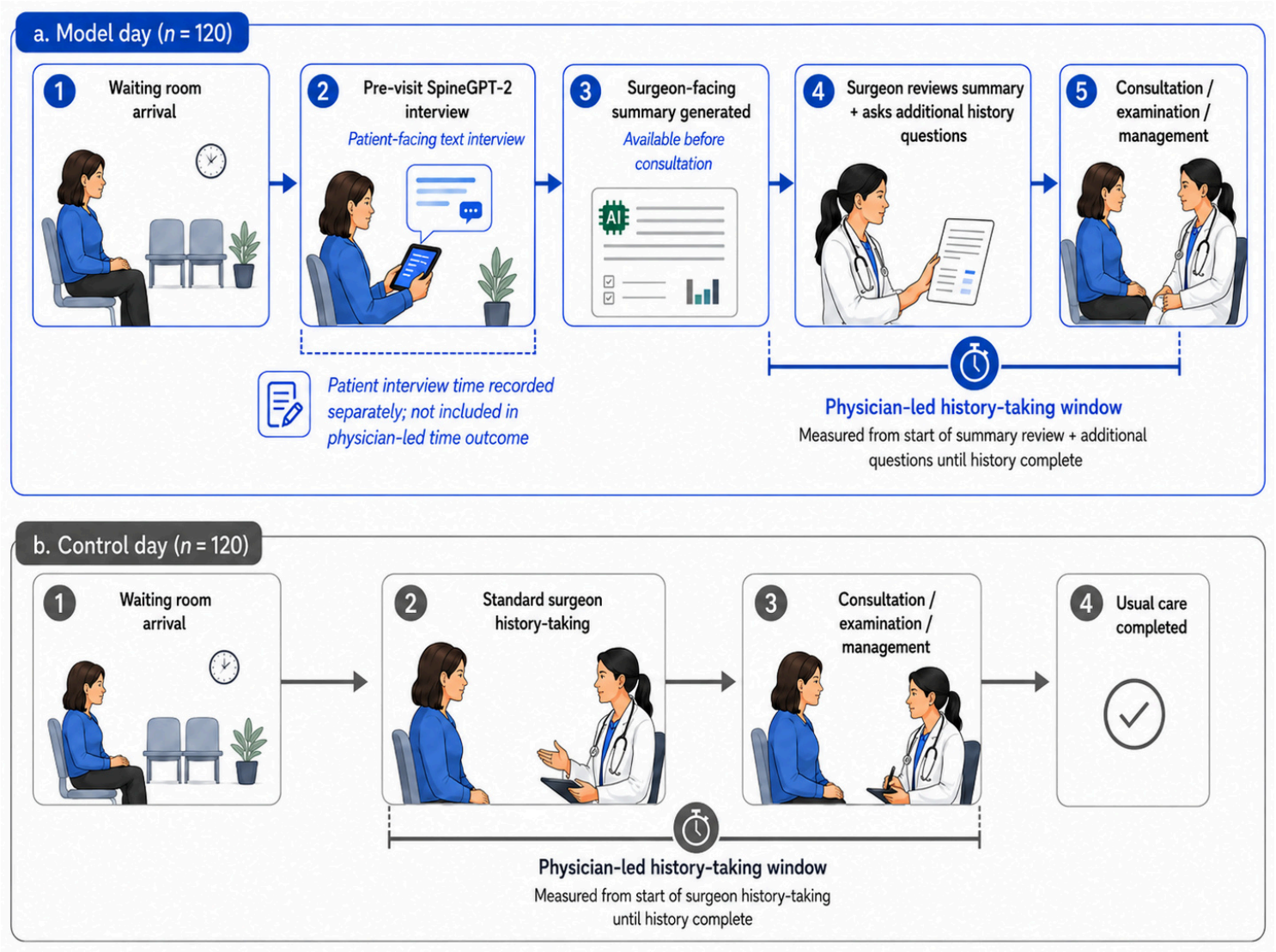


Figure 7

Clinic days alternated between the model and control pathways, yielding 15 clinic days and 120 patients in each arm. a. On model days, patients completed a pre-visit SpineGPT-2 interview in the waiting area, after which the surgeon-facing summary was generated and made available before consultation. The physician-led history-taking window began when the surgeon reviewed the summary and asked any additional history questions and ended when history-taking was complete. The patient-facing interview time was recorded separately and excluded from the consultation-time comparison. b. On control days, the physician-led history-taking window began when the surgeon started standard history-taking and ended when the history was complete. Examination and clinical management subsequently proceeded according to usual care in both arms. The numbered blocks represent clinic-day sequence rather than calendar dates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf