SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 210207:1–210207:10
https://doi.org/10.1007/s11432-026-5234-1

Special Topic: Intelligent Characterization and Modeling for Complex Systems

Clinical implementation and evaluation of an LLM-based conversational agent for preanesthetic interviews

Hong TAO[1†], Yi-Fan HU[4†], Hui YU[2†], Chun (Martin) QIU[3], Yi-Fan HAN[4], Pei LU[1], Cheng CHENG[4], Zhi-Qiang ZHOU[1], Zhi-Quan LI[4], Hao CHEN[1], Yue-Yang XIN[1], Li-Jun ZHU[4], Han DING[5*], Hui XU[1*] & Hai-Tao ZHANG[4*]

[1]*Department of Anesthesiology and Pain Medicine, Hubei Key Laboratory of Geriatric Anesthesia and Perioperative Brain Health, Wuhan Clinical Research Center for Geriatric Anesthesia, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China*
[2]*Wuhan Children's Hospital (Wuhan Maternal and Child Healthcare Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430010, China*
[3]*Lazaridis School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada*
[4]*School of Artificial Intelligence and Automation, the MOE Engineering Research Center of Autonomous Intelligent Unmanned Systems and State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[5]*State Key Laboratory of Intelligent Manufacturing Equipment and Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*

Received 8 June 2026/Revised 9 October 2026/Accepted 16 October 2026/Published online 19 November 2026

Abstract Preanesthetic interviews are essential for risk assessment and patient education, but surgical volume growth and anesthesiology workforce shortages constrain consultation time. To this end, we built a system combining multi-agent orchestration, knowledge-graph retrieval, and local LLM deployment, and accordingly developed Shuwen, an LLM-based conversational agent that conducts structured interviews, provides perioperative counseling, and generates draft documentation for anesthesiologist verification. In this prospective single-center study involving 332 adult elective surgery patients, the primary outcome was document-level accuracy of AI-generated summaries (composite of medications, diagnoses, and allergies), assessed by majority rule ($\geqslant$ 2 of 3 anesthesiologist raters) against clinician-verified records. Shuwen completed 321 interviews without mid-session takeover (96.7%). Document-level accuracy was 84.9% (95% CI, 80.7%–88.4%), with category-level accuracies of 96.4% for medications, 90.4% for diagnoses, and 97.3% for allergies. Accuracy declined with higher ASA status (ASA I: 92.6%; ASA II: 84.6%; ASA III: 74.2%), driven primarily by decreasing diagnostic accuracy. Median interview duration was 11.3 min. Patient satisfaction exceeded the satisfactory benchmark across all domains (mean range, 4.5–4.8). Anesthesiologists have reported favorable evaluations (mean range, 4.2–4.8). In a predeployment assessment of 315 adults, 46.3% were unable to distinguish Shuwen's responses from those of anesthesiologists. These findings support the feasibility of LLM-mediated preanesthetic interviewing; effects on workload, efficiency, and care quality require controlled evaluation. From a broader perspective, the results shed light on how leveraging recent advances in agentic LLMs and knowledge-grounded AI helps to reveal latent workflow structure. Moreover, they help pave the way for converting complex clinical interactions into interpretable, data-driven decision support in safety-critical environments.

Keywords multi-agent systems, complex networks, large language model, network control, clinical AI

Citation Tao H, Hu Y-F, Yu H, et al. Clinical implementation and evaluation of an LLM-based conversational agent for preanesthetic interviews. Sci China Inf Sci, 2026, 69(11): 210207, https://doi.org/10.1007/s11432-026-5234-1

1 Introduction

Preanesthetic interviews are a core element of international perioperative care standards, enabling anesthesiologists to assess risks, optimize comorbid conditions, review pertinent data, and communicate anesthesia plans [1,2]. These encounters can also alleviate anxiety and support recovery [2]. Yet rising surgical demand and limited anesthesiology staffing have constrained the time available for patient engagement in many health systems, including those in the United States and China. In the United States, a persistent and widening shortage of anesthesiologists has been reported, with the deficit projected to reach 2.6%–12% annually given the 2%–3% yearly increase in surgical volume [3]. In China, anesthesiologists frequently work more than 60 h per week, and the workforce density is

* Corresponding author (email: dinghan@hust.edu.cn, huixu@tjh.tjmu.edu.cn, zht@mail.hust.edu.cn)
† These authors contributed equally to this work.

© The Author(s) 2026. This article is published with open access at link.springer.com.

6.53 per 100000 population, which is below the global average of 7.60 and lower than levels observed in countries with comparable economic development [4–6]. Short visit lengths and competing workload pressures may leave patients insufficiently informed about anesthesia-related risks and without adequate reassurance. In clinical practice, a thorough preanesthetic interview conducted by anesthesiologists typically requires 18–22 min on average [7, 8]. This duration is known to increase with patient complexity, such as higher ASA physical status, due to the need for more detailed risk discussion and management planning [7].

Large language model (LLM)-based tools offer a potential approach to scale patient communication while preserving clinical relevance. Recent work on LLM-based agents and knowledge-enhanced domain-specific AI further supports the use of agentic task decomposition and domain knowledge grounding for specialized workflows [9, 10]. More broadly, these advances reflect an emerging AI-enabled paradigm for extracting structured variables, identifying hidden workflow states, and generating compact, interpretable representations from complex clinical data. Multimodal LLMs have been applied across a variety of medical domains, including clinical decision support [11,12], automated medical documentation [13], patient education, and biomedical research [14]. Generative AI voice agents are increasingly recognized as transformative technologies for medical applications [15,16]. AI applications in anesthesiology have expanded across clinical workflow, from preoperative assessment to intraoperative support [17]. Early prototypes have demonstrated promise for explanation and anxiety reduction in simulated scenarios [18]; however, relatively few reported systems have addressed the full scope of real-world patient interaction within preanesthetic evaluation, including structured history taking, patient education, draft documentation, and clinician sign-off [19, 20].

As a remedy, we built Shuwen, an LLM-powered conversational agent designed to function as both interviewer and provider of preoperative patient education. Distinct from prior systems focused primarily on documentation, Shuwen conducts structured history taking, explains perioperative risks and multimodal analgesia, responds to patient questions with attention to clarity and empathy, and generates draft documentation. A multi-agent orchestration framework, grounded in a perioperative knowledge graph, is established with the design objective of supporting safety, consistency, and traceability in dialogue [20]. From the perspective of complex system modeling, the present clinical workflow constitutes a nonlinear, multivariate process with interdependent information sources and cascading safety constraints; the knowledge-graph–augmented multi-agent framework can be viewed as a principled approach to decomposing this complexity into coordinated, interpretable sub-tasks. In this way, the present design offers a generalizable paradigm for revealing hidden workflow states and managing complexity in healthcare delivery systems.

This study evaluates the clinical performance of Shuwen through a prospective observational study involving 332 surgical patients and 11 anesthesiologists in a local hospital. We assessed documentation accuracy, patient experience, and clinician acceptance in a high-volume academic medical center. Patient perspectives on AI in healthcare have emphasized the importance of transparency, accountability, and maintaining human oversight [21, 22]. The results demonstrate the practical feasibility of an LLM-based interview system supported by a perioperative knowledge graph within a patient-centered preanesthetic workflow. Shuwen achieved a 96.7% interview-completion rate without mid-session takeover, with 46.3% of participants unable to distinguish its responses from those of anesthesiologists in a blinded Turing-style assessment. This human-likeness endpoint reflects perceived conversational similarity and is not a measure of clinical safety or completeness.

2 Problem formulation

2.1 Study design and participants

This prospective, single-center observational implementation study was conducted at a tertiary academic medical center from April 1 to August 30, 2025. The study protocol was approved by the local Institutional Review Board (Approval No. [2025] S024), and written informed consent was obtained from all participants. The study adhered to STROBE guidelines and CHART considerations for studies involving conversational AI [23].

Consecutive adults scheduled for elective surgery in general surgery, thoracic surgery, orthopedics, gynecology, urology, and other specialties were enrolled. Exclusion criteria were cognitive impairment precluding informed consent or insufficient language proficiency to complete the interview. Demographic and clinical characteristics were extracted from the electronic health record (EHR). The target sample size of approximately 330 interviews was determined according to the expected clinic volume to ensure adequate precision of accuracy estimates. With $N = 332$, the two-sided 95% CI half-width for a proportion of 0.90 is approximately $\pm$ 3.4 percentage points, and for 0.95 is approximately $\pm$ 2.5 percentage points (Clopper-Pearson intervals).

2.2 Outcome measures and statistical analysis

The primary outcome was document-level accuracy of AI-generated preanesthetic summaries, defined as a composite binary measure indicating whether medications, diagnoses, and allergies were all correct. These domains were selected for their safety relevance and consistent documentation across patients. Prior anesthesia history, airway information, fasting status, and preferences were collected but excluded from the composite because their applicability and documentation varied by patient. Three senior anesthesiologists, each with over 10 years of clinical experience, independently assessed accuracy across three prespecified categories using a binary scale (correct vs. incorrect). Raters were blinded to each other's assessments and followed standardized instructions. All three raters participated in clinical preanesthetic encounters and final-record review or sign-off, and contributed to the structured portion of the perioperative knowledge graph. Accuracy was determined relative to the anesthesiologist-verified final chart as the reference standard. A document was considered correct by a given rater only if all three categories were rated as "correct". Disagreements among anesthesiologists were not adjudicated; instead, accuracy for each category and the overall document was determined by majority rule ($\geqslant$ 2 of 3 raters). The final chart incorporated review of the AI draft; accuracy therefore represents concordance with the clinician-verified record.

Interrater agreement was quantified using Gwet's AC1 statistic with 95% confidence intervals (CIs), estimated for all three anesthesiologists jointly and via leave-one-out pairwise sensitivity analyses. The Gwet's AC1 statistic is defined as

$$\mathrm{AC1} = \frac{p_a - p_e}{1 - p_e}, \tag{1}$$

where p_a is the observed agreement and p_e is the chance agreement. For N subjects and 3 raters, the observed agreement p_a is calculated as

$$p_a = \frac{1}{N}\sum_{i=1}^{N}\frac{n_i(n_i-1)+(3-n_i)(3-n_i-1)}{6}, \tag{2}$$

where n_i represents the number of raters who classified subject i as "correct" (or the category of interest). The chance agreement p_e is determined by the prevalence probability π:

$$p_e = 2\pi(1-\pi). \tag{3}$$

The prevalence π is estimated by

$$\pi = \frac{1}{3N}\sum_{i=1}^{N} Y_i, \tag{4}$$

where Y_i represents the total number of "YES" votes for subject i. These calculations follow the methodology described in [24]. Accuracy proportions are reported with 95% Clopper-Pearson CIs for category-level and document-level assessments. All tests were two-sided with $\alpha = 0.05$. Statistical analyses were conducted using R version 4.3.2 (irrCAC package). Reproducible code is available upon request (Appendix H).

For patient and clinician experience, domain scores are presented as mean (SD) and median (IQR). One-sample Wilcoxon signed-rank tests compared domain medians against the prespecified satisfactory reference of 4.0. Independent samples t-tests compared satisfaction scores between Shuwen and human doctor groups across the five domains, including interaction efficiency, information clarity, professional competence, empathy, and overall satisfaction. Raw P values are reported in the tables, and Holm-adjusted results were used for multiplicity-sensitive interpretation across the 5 patient and 8 clinician domains. CFA used WLSMV estimation on polychoric correlations; standardized factor loadings, fit indices, and Cronbach's α/composite reliability are reported. Analyses were conducted in R version 4.3.2 and Python version 3.11 using packages irrCAC, psych, lavaan, and statsmodels/scipy (Appendixes F and G).

3 System design

3.1 System architecture

Shuwen is an LLM-driven conversational AI agent designed for preanesthetic interviews. Figure 1 summarizes the system architecture and workflow. The AI-agent-based system is composed of three components: (1) a multi-agent orchestration framework that delegates clinical sub-tasks to role-constrained agents and is designed to reduce

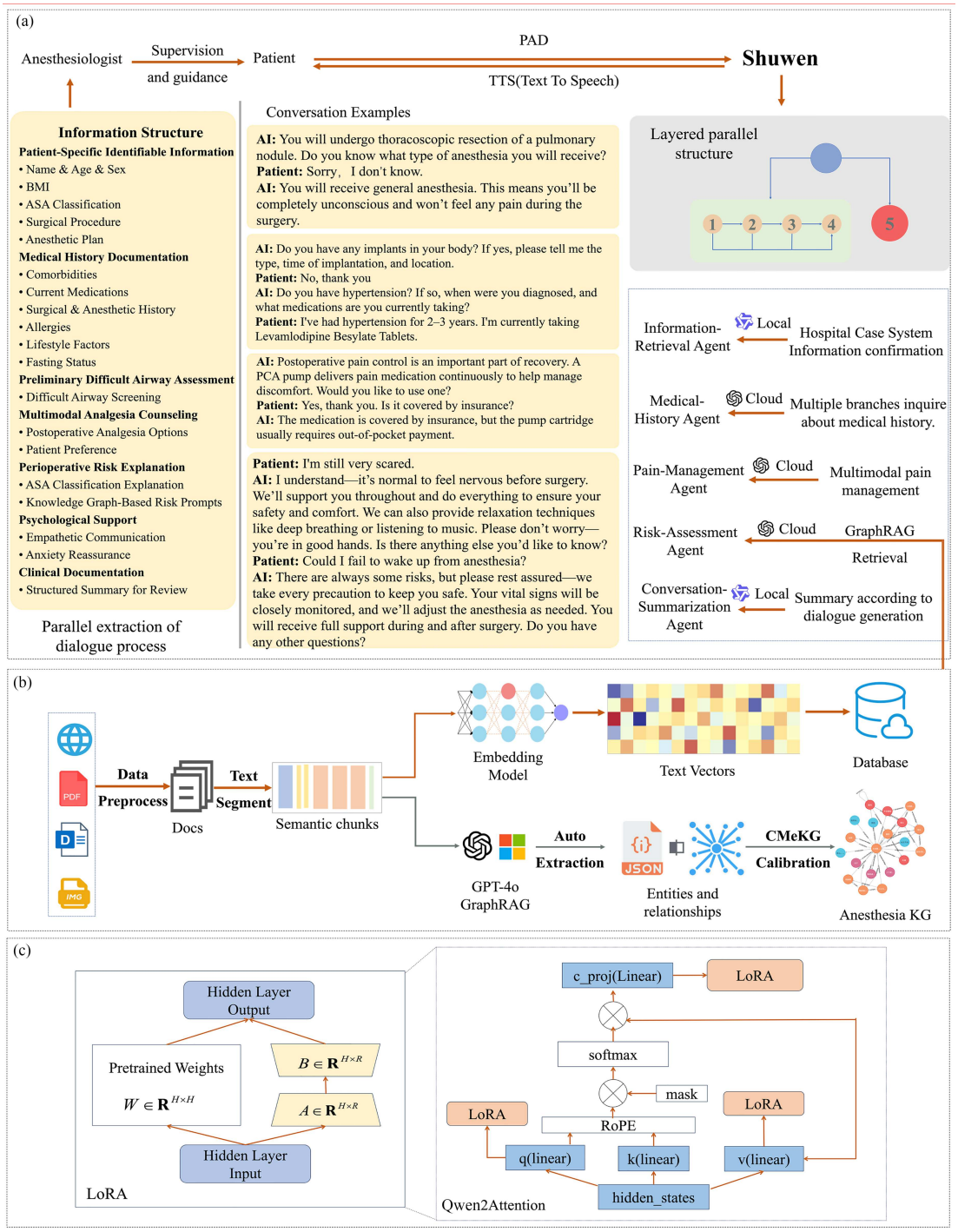

Figure 1 (Color online) Architecture and workflow of the AI-based preanesthetic interview system (Shuwen), organized into three visual layers. (a) Multi-agent collaborative and clinical workflow. The system is orchestrated by a central task-scheduling agent that coordinates five specialized functional agents: (1) an information-retrieval agent that interfaces with the hospital electronic medical record to confirm patient identity and demographics; (2) a medical-history agent that conducts structured, interactive history taking; (3) a pain-management agent that discusses multimodal postoperative analgesia options; (4) a risk-assessment agent that applies graph-enhanced retrieval and generation to evaluate individualized perioperative risk; and (5) a conversation-summarization agent that compiles structured clinical summaries for anesthesiologist review. The left panel lists the predefined information structure captured during the interview, with representative patient-AI dialogue excerpts shown in the center. (b) Knowledge-graph construction and GraphRAG workflow. Clinical documents in multiple formats (e.g., PDF, Docs, image) are preprocessed and segmented into semantic units, converted into vector representations via an embedding model, and stored in a dedicated database. Entity and relationship extraction is performed using GPT-4o with GraphRAG, and results are cross-validated against the Chinese medical knowledge graph [29] to construct an anesthesia-specific knowledge graph. This curated resource supports contextually grounded and auditable outputs during risk assessment. (c) LoRA fine-tuning implementation for domain adaptation of the locally deployed Qwen model.

task drift and enhance fault tolerance [25]; (2) a perioperative knowledge graph coupled with graph retrieval-augmented generation (GraphRAG) to anchor responses to guideline-linked, auditable sources [26]; and (3) a real-time voice/text interface with layered safety controls. By making the retrieval source explicit, the knowledge graph is intended to support factuality control by injecting domain knowledge into generation while constraining responses to clinically auditable relationships, a direction consistent with recent work on knowledge injection for LLMs [27]. Explainable AI principles are increasingly important for healthcare applications to ensure transparency and trust in clinical decision support [28]. Ethical considerations surrounding the deployment of LLMs in medical settings, including issues of bias, accountability, and patient safety, have been extensively discussed in the literature.

The agents are prompt-defined roles sharing patient state. A deterministic scheduler follows the clinical stage sequence, with patient-dependent follow-up within each stage. Shared state maps dialogue stages and required or optional information slots to permitted agent actions, retrieved evidence, and structured notes. Missing or uncertain fields remain unresolved for clinician verification. This mapping makes the otherwise implicit clinical workflow explicit.

Knowledge assets were curated from standardized perioperative forms, pharmacological guidance, and procedure-specific protocols, then cross-validated against the Chinese medical knowledge graph (CMeKG) through entity alignment, semantic consistency checking, and manual review of ambiguous concepts [29]. During risk assessment and medication guidance, vector retrieval combined with graph indexing supplied five candidate items; corresponding triples were inserted into the active role prompt with patient state. Graph scale and types are detailed in Appendix B.3. To meet institutional privacy requirements, protected health information was processed on-premises using a fixed local Qwen-2.5-14B checkpoint. The model was adapted to perioperative dialogue through LoRA fine-tuning. External processing was limited to de-identified text, with identifiers removed primarily by automated rules and limited manual review by study researchers. The Qwen-2.5-14B model is hosted on a server with four NVIDIA 4090 GPUs. Safety measures include role-constrained prompting, knowledge-graph scoping, deterministic fallbacks for safety-critical content, automated escalation triggers, and operational dialogue logging.

3.2 Clinical implementation

Each session began with identity verification and demographic confirmation against the EHR, followed by the Shuwen interview on a tablet interface supporting low-latency speech synthesis (iFlytek) and text input/output. The interview system elicited comorbidities, prior surgical and anesthetic history, allergies, medications, lifestyle factors, airway risk, and fasting status. The agent also delivered individualized, graph-anchored perioperative counseling while documenting patient preferences for multimodal analgesia. Prompt templates specified the clinical role, scope restrictions, required information slots, safety instructions, one-question-at-a-time elicitation, explicit handling of missing information, and completion or clinician-handoff conditions. Response latency and failure handling are described in Appendix B.5. The prompting strategy for logical questioning is provided in Appendix C.

Upon completion, an anesthesiologist reviewed the AI-generated structured note, performed a targeted physical examination, verified critical information, and further explored predefined high-risk factors (e.g., allergies, active chest pain, syncope). Discrepancies were resolved before sign-off and submission to the medical record.

3.3 Data collection

Before clinical deployment, a blinded Turing-style evaluation was conducted in an independent cohort of 315 adults. Participants reviewed approximately 40–50 paired responses, with the exact number varying according to patient-specific follow-up needs, across history taking, risk explanation, medication guidance, pain management, and general communication. Each pair contained one response authored by Shuwen and one extracted from a real anesthesiologist-patient interaction. The source was concealed, the left-right presentation order was randomized, and neither participants nor study personnel were informed of the response source during assessment. Participants included adults of both sexes and varied ages and backgrounds, with only a small minority having medical training. The indistinguishability rate (proportion failing to identify the clinician response) was prespecified as the endpoint. This assessment measured perceived conversational similarity under the sampled questions and did not assess factual completeness or clinical safety.

Immediately after the AI interview and before anesthesiologist sign-off, patients completed a 14-item questionnaire covering five domains (overall satisfaction, information clarity, interaction efficiency, empathy, and perceived professional competence) on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). For comparison, satisfaction ratings were also collected from patients who received traditional preanesthetic interviews conducted by human anesthesiologists using the same instrument. Eleven anesthesiologists (2 chief, 2 associate chief, 3 attending,

Table 1 Demographics of study participants. Data are mean (SD) or n (%). ASA, American Society of Anesthesiologists. "No prior exposure" denotes patients without previous general/regional anesthesia.

Characteristic	Total (n = 332)	Characteristic	Total (n = 332)
Age, years (mean [SD])	50.3 (12.4)	Education level, n (%)	
Sex, n (%)		Basic	164 (49.4)
Women	168 (50.6)	Secondary	61 (18.4)
Men	164 (49.4)	Higher	107 (32.2)
Age group, n (%)		Anesthesia history, n (%)	
19–39	76 (22.9)	Prior exposure	242 (72.9)
40–59	169 (50.9)	No prior exposure	90 (27.1)
60–80	87 (26.2)		
Surgical specialty, n (%)		ASA physical status, n (%)	
General abdominal	194 (58.4)	I	54 (16.3)
Thoracic	59 (17.8)	II	247 (74.4)
Urologic	42 (12.7)	III	31 (9.3)
Orthopedic	30 (9.0)		
Gynecological	3 (0.9)		
Other	4 (1.2)		

4 residents) completed an 8-domain evaluation using the same scale. Item stems and anchors are provided in Appendix E. The survey structure was adapted from the Patient Satisfaction Questionnaire Short Form (PSQ-18) and tailored to perioperative AI use [30]. Confirmatory factor analysis (CFA) evaluated the latent structure and reliability. Model fit was assessed using CFI/TLI, RMSEA, and SRMR (target thresholds: CFI/TLI $\geqslant 0.95$; RMSEA $\leqslant 0.08$; SRMR $\leqslant 0.08$). Internal consistency was summarized with Cronbach's α and composite reliability.

4 Main technical results

4.1 Patient demographics

Of 338 patients invited, six sessions were not completed (four patients chose to speak directly with a physician; two interviews were interrupted due to network failures). All six transferred to the physician-led pathway; the small number precluded inference of a common pattern. The final cohort comprised 332 patients with a mean (SD) age of 50.3 (12.4) years and balanced sex distribution (168 women [50.6%], 164 men [49.4%]). The predominant surgical category was general abdominal (58.4%), followed by thoracic (17.8%), urologic (12.7%), orthopedic (9.0%), gynecological (0.9%), and other procedures (1.2%). Most patients were classified as ASA II (74.4%), with ASA I (16.3%) and ASA III (9.3%). Elderly patients with limited reading ability received caregiver support as needed; fewer than 5% required brief initial assistance (e.g., initiating the session) before continuing independently (Table 1). This assistance may have influenced experience ratings.

4.2 Accuracy of Shuwen-generated documentation

Accuracy was assessed across three predefined categories (medications, diagnoses, and allergies) based on interrater agreement among three anesthesiologists. Category-level accuracies were 96.4% for medications (320/332; 95% CI, 93.8%–97.9%), 90.4% for diagnoses (300/332; 95% CI, 86.7%–93.1%), and 97.3% for allergies (323/332; 95% CI, 94.9%–98.6%). Overall document-level accuracy was 84.9% (282/332; 95% CI, 80.7%–88.4%; Table 2). Category-level accuracy was lowest for diagnoses, consistent with prior work showing that structured medication and allergy information can be extracted more reliably than diagnostic labels from clinical text and speech [31].

Interrater agreement was quantified using Gwet's AC1 statistic. Agreement was high across all three categories (three-rater AC1 range, 0.9–1.0; all $P < 0.001$). Leave-one-out pairwise sensitivity analyses demonstrated strong agreement (AC1 range, 0.9–1.0; all $P < 0.001$).

4.3 Patient-reported experience

The median Shuwen interview duration was 11.3 min (IQR 9.4–13.6), with longer durations for patients with complex medical histories. Follow-up anesthesiologist consultations averaged approximately five additional minutes.

Table 2 Performance metrics of the Shuwen preanesthetic interview system. * Accuracy was determined by simple majority vote (at least two of the three reviewers agreed the note was accurate). ** Document-level accuracy required all three categories to be rated accurate by an anesthesiologist rater; consensus was then determined by majority rule (⩾2/3). *** Duration is reported as median (IQR) because of non-normal distribution. Proportions are shown with 95% CIs (Clopper-Pearson).

Metric	Overall	ASA I	ASA II	ASA III
Interview completion without mid-session takeover	321/332; 96.7 (94.2–98.1)	–	–	–
Accuracy, n/total; % (95% CI)				
Medication list*	320/332; 96.4 (93.8–97.9)	54/54; 100.0 (93.4–100.0)	238/247; 96.4 (93.2–98.1)	28/31; 90.3 (75.1–96.7)
Key diagnoses*	300/332; 90.4 (86.7–93.1)	52/54; 96.3 (87.5–99.0)	222/247; 89.9 (85.5–93.1)	26/31; 83.9 (67.4–92.9)
Allergy list*	323/332; 97.3 (94.9–98.6)	52/54; 96.3 (87.5–99.0)	240/247; 97.2 (94.3–98.6)	31/31; 100.0 (89.0–100.0)
Overall accuracy**	282/332; 84.9 (80.7–88.4)	50/54; 92.6 (82.5–97.1)	209/247; 84.6 (79.6–88.6)	23/31; 74.2 (56.8–86.3)
Conversation duration, median (IQR), min***	11.3 (9.4–13.6)	–	–	–

Table 3 Comparison of patient satisfaction scores between Shuwen (AI) and human doctors. Scores are mean ± SD. *** indicates statistical significance ($P < 0.001$).

Dimension	Shuwen	Human doctors	t statistic	P value
Overall satisfaction	4.55 ± 0.50	4.58 ± 0.49	−0.924	0.356
Interaction efficiency	4.50 ± 0.50	4.57 ± 0.50	−1.706	0.089
Information clarity	4.53 ± 0.50	4.56 ± 0.50	−0.773	0.440
Empathy	4.76 ± 0.43	4.60 ± 0.49	4.436	$< 0.001^{***}$
Professional competence	4.55 ± 0.50	4.53 ± 0.50	0.456	0.648

Confirmatory factor analysis supported the predefined five-domain structure (Comparative Fit Index 0.96, Tucker-Lewis Index 0.95, RMSEA 0.08, SRMR 0.03). Internal consistency was high across domains (Cronbach's α 0.84–0.89).

Mean domain scores for both Shuwen and human doctors exceeded the satisfactory benchmark of 4.0. Shuwen achieved significantly higher empathy ratings (4.8 ± 0.4 vs. 4.6 ± 0.5; $P < 0.001$; Table 3). No significant differences were observed in overall satisfaction (4.5 ± 0.5 vs. 4.6 ± 0.5; $P = 0.356$), interaction efficiency (4.5 ± 0.5 vs. 4.6 ± 0.5; $P = 0.089$), information clarity (4.5 ± 0.5 vs. 4.6 ± 0.5; $P = 0.440$), or professional competence (4.5 ± 0.5 vs. 4.5 ± 0.5; $P = 0.648$). This finding aligns with recent studies demonstrating that patients often perceive AI responses as empathetic as or more empathetic than physician responses [32]. Figure 2(a) shows Shuwen achieved superior performance in empathy while maintaining parity in other dimensions, indicating that structured, LLM-guided interviews can maintain patient-centered communication [33]. Domain scores did not differ significantly by sex, age group, or education level (Figures 2(b)–(d)). Overall, 135 patients (40.7%) expressed a preference for AI-assisted preoperative visits. Acceptance of AI-mediated communication has been observed across various telemedicine and digital health platforms, suggesting growing patient comfort with technology-assisted care [34]. High satisfaction thus coexisted with a minority preference for AI involvement.

4.4 Anesthesiologist evaluations

Eleven anesthesiologists evaluated Shuwen across eight domains. Mean (SD) scores were: overall satisfaction 4.7 (0.5), accuracy 4.7 (0.5), safety 4.6 (0.4), communication 4.8 (0.4), practicality 4.5 (0.5), efficiency 4.3 (0.6), accessibility 4.2 (0.5), and preference for continued use 4.6 (0.5). These ratings indicate clinician confidence in the system's reliability and communication quality. Professional perspectives on healthcare AI have highlighted both opportunities for workflow enhancement and concerns about changes in clinical roles and skills [35]. Slightly lower scores for practicality and efficiency pointed to residual frictions that could be addressed through workflow optimization, consistent with reports that ambient documentation tools are associated with reduced documentation burden and clinician burnout [36, 37]. Proper informed consent processes for AI-assisted documentation remain an important consideration for clinical deployment [38]. These ratings represent clinician perceptions and do not directly quantify workload reduction or clinical efficiency.

4.5 Predeployment human-likeness assessment

Prior to clinical deployment, a blinded Turing test was conducted with an independent cohort of 315 participants. The indistinguishability rate was 46.3% (146 participants unable to distinguish Shuwen's responses from those of

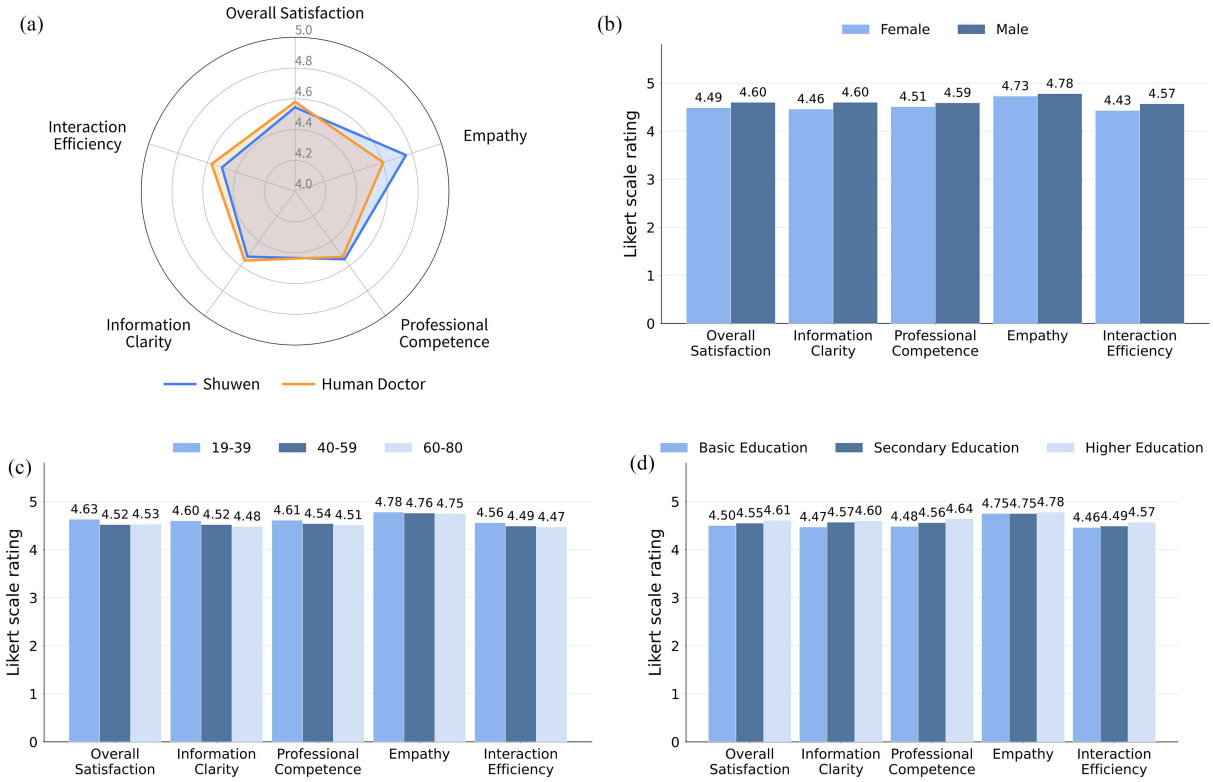

Figure 2 (Color online) Patient-reported satisfaction across five dimensions. (a) Radar chart comparing Shuwen and human doctors, showing high and balanced ratings for both groups, with Shuwen achieving significantly higher scores in empathy. (b)–(d) Stratified ratings by sex, age group, and education level, showing consistent performance across demographics. While most dimensions showed no significant differences (all $P > 0.05$), Shuwen consistently outperformed human doctors in empathy across multiple subgroups ($P < 0.05$).

anesthesiologists). This result concerns perceived authorship and should not be interpreted as evidence of clinical safety, completeness, or equivalence.

5 Discussion

Shuwen's system-level contribution is the shared-state coordination of history taking, patient education, knowledge-grounded risk assessment, and clinician-verified documentation. Role and scope constraints support response consistency, graph retrieval supports factuality control, and operational logs support traceability. Required fields and stage transitions define interview completion, while unresolved information is preserved for clinician review. These mechanisms connect the information architecture to the clinical workflow.

High satisfaction alongside a 40.7% preference for AI assistance suggests that patients may value clarity and convenience while continuing to prefer physician involvement in perioperative decisions. Patient choice and technical interruption therefore remain integral to the handoff pathway.

The single-center observational design lacked a concurrent clinician-only control for documentation or workflow outcomes and component-level ablation. Clinical findings characterize the implemented system; controlled comparisons are needed to establish workload, efficiency, and clinical benefits or individual component contributions. The accuracy composite covers selected documentation domains, and the final record's dependence on AI-draft review and the raters' clinical and development roles may introduce bias. Human-likeness reflects perceived authorship rather than clinical reliability. Multicenter evaluation should include local protocols, terminology, knowledge assets, workflow coordination, and governance.

6 Conclusion

This study developed and evaluated Shuwen, an AI-agent-based conversational LLM system for preanesthetic history-taking and perioperative education, followed by anesthesiologist confirmation within the clinical workflow. The findings support its practical value for structured information acquisition and perioperative education within a

clinician-supervised workflow. The system achieved a high interview-completion rate without mid-session takeover (96.7%) and document-level accuracy (84.9%), with favorable patient and clinician acceptance. Beyond the immediate clinical setting, this work exemplifies how recent advances in agentic LLMs and knowledge-grounded AI can transform high-dimensional, interdependent clinical processes into structured, interpretable system representations, offering a generalizable paradigm for revealing hidden workflow states and managing complexity in healthcare delivery systems.

Future work includes controlled baseline and component-ablation studies and multicenter prospective cohort studies to validate these findings across diverse institutions and patient populations, and to assess impacts on surgical cancellations, complications, pain control, and perioperative anxiety. Establishing AI implementation science centers has been proposed as a critical infrastructure for rigorous evaluation of clinical effectiveness and real-world performance of AI systems [39]. Technical development focuses on seamless EHR connectivity, adaptive questioning for complex cases, and expansion of curated medical knowledge. While Shuwen was developed in Chinese clinical settings, its core approach of LLM-mediated interviewing grounded in a curated knowledge graph can be adapted to other languages and health systems through localization of knowledge assets and workflows.

Acknowledgements Hai-Tao ZHANG's work was supported in part by National Natural Science Foundation of China for Distinguished Young Scholars (Grant No. 62225306). Hui XU's work was supported in part by High-Quality Clinical Research Fund of Tongji Hospital (Grant No. 2024TJCR025). This work was supported by Interdisciplinary Research Program of HUST (Grant No. 2024JCYJ033) and National Natural Science Foundation of China (Grant No. 52188102).

Supporting information Appendixes A–H. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the articles Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1 Harms C, Kindler C H. The preoperative anaesthetic visit. Ther Umschau, 2009, 66: 503–508
2 Jindal P, Patil V, Pradhan R, et al. Update on preoperative evaluation and optimisation. Ind J Anaesthesia, 2023, 67: 39–47
3 Menezes J, Zahalka C. Anesthesiologist shortage in the United States: a call for action. J Med Surg Public Health, 2024, 2: 100048
4 Chong M Y F, Lin S H X, Lim W Y, et al. Burnout in anaesthesiology residents: a systematic review of its prevalence and stressors. Eur J Anaesth, 2022, 39: 368–377
5 Li H, Zuo M, Gelb A W, et al. Chinese anesthesiologists have high burnout and low job satisfaction: a cross-sectional survey. Anesth Analg, 2018, 126: 1004–1012
6 Law T J, Lipnick M S, Morriss W, et al. The global anesthesia workforce survey: updates and trends in the anesthesia workforce. Anesth Analg, 2024, 139: 15–24
7 Lassen C L, Jaschinsky F, Stamouli E, et al. Anesthesiological preoperative interview with a palliative care patient: a simulation-based experiment using standardized patients. Medicina, 2024, 60: 1577
8 Salzwedel C, Petersen C, Blanc I, et al. The effect of detailed, video-assisted anesthesia risk education on patient anxiety and the duration of the preanesthetic interview: a randomized controlled trial. Anesth Analg, 2008, 106: 202–209
9 Xi Z H, Chen W X, Guo X, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci, 2025, 68: 121101
10 Li X Y, Wang Y, Liang J, et al. Knowledge-enhanced AI models for domain-specific application: a survey. Sci China Inf Sci, 2026, 69: 141101
11 Xiong H, Zheng Z, Qiu Z P, et al. Prescription generation via multi-source context-aware hierarchical transformer. Sci China Inf Sci, 2026, 69: 122109
12 Li J, Deng Y, Sun Q, et al. Benchmarking large language models in evidence-based medicine. IEEE J Biomed Health Inform, 2025, 29: 6143–6156
13 Feng H, Liu Q, Liu H, et al. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. Sci China Inf Sci, 2024, 67: 220106
14 Luo Y, Shi L, Li Y, et al. From intention to implementation: automating biomedical research via LLMs. Sci China Inf Sci, 2025, 68: 170105
15 Adams S J, Acosta J N, Rajpurkar P. How generative AI voice agents will transform medicine. npj Digit Med, 2025, 8: 353
16 Tomczak J M. Generative AI systems: a systems-based perspective on generative AI. 2024. ArXiv:2407.11001

17 Hashimoto D A, Witkowski E, Gao L, et al. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology, 2020, 132: 379–394

18 Yahagi M, Hiruta R, Miyauchi C, et al. Comparison of conventional anesthesia nurse education and an artificial intelligence chatbot (ChatGPT) intervention on preoperative anxiety: a randomized controlled trial. J PeriAnesth Nurs, 2024, 39: 767–771

19 McPeak G, Sautmann A, George O, et al. An LLM's medical testing recommendations in a Nigerian clinic: potential and limits of prompt engineering for clinical decision support. In: Proceedings of the 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), 2024. 586–591

20 Crema C, Verde F, Tiraboschi P, et al. Medical information extraction with NLP-powered QABots: a real-world scenario. IEEE J Biomed Health Inform, 2024, 28: 6906–6917

21 Khullar D, Casalino L P, Qian Y, et al. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open, 2022, 5: e2210309

22 Ellis J, Hamer M K, Akerson M, et al. Patient perceptions of chatbot supervision in health care settings. JAMA Netw Open, 2024, 7: e248833

23 von Elm E, Altman D G, Egger M, et al. The Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet, 2007, 370: 1453–1457

24 Gwet K L. Computing inter-rater reliability and its variance in the presence of high agreement. Brit J Math Statis, 2008, 61: 29–48

25 Chen X, Yi H, You M, et al. Enhancing diagnostic capability with multi-agents conversational large language models. npj Digit Med, 2025, 8: 159

26 Guo Z, Xia L, Yu Y, et al. LightRAG: simple and fast retrieval-augmented generation. 2024. ArXiv:2410.05779

27 Ke W, Shang Z, Luo Z, et al. Unveiling factuality and injecting knowledge for LLMs via reinforcement learning and data proportion. Sci China Inf Sci, 2024, 67: 209101

28 Bharati S, Mondal M R H, Podder P. A review on explainable artificial intelligence for healthcare: why, how, and when? IEEE Trans Artif Intell, 2024, 5: 1429–1442

29 Li X, Yuan Y, Yang Y, et al. Quality-controllable automatic construction method of Chinese knowledge graph for medical decision-making applications. Inf Process Manag, 2025, 62: 104148

30 Thayaparan A J, Mahdi E. The patient satisfaction questionnaire short form (PSQ-18) as an adaptable, reliable, and validated tool for use in various settings. Med Educ Online, 2013, 18: 21747

31 Omiye J A, Gui H, Rezaei S J, et al. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med, 2024, 177: 210–220

32 Chen D, Chauhan K, Parsa R, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. npj Digit Med, 2025, 8: 275

33 Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digit Med, 2024, 7: 82

34 Harnik M A, Scheidegger A, Blättler L, et al. Acceptance, satisfaction, and preference with telemedicine during the COVID-19 pandemic in 2021-2022: survey among patients with chronic pain. JMIR Form Res, 2024, 8: e53154

35 Aquino Y S J, Rogers W A, Braunack-Mayer A, et al. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inf, 2023, 169: 104903

36 You J G, Dbouk R H, Landman A, et al. Ambient documentation technology in clinician experience of documentation burden and burnout. JAMA Netw Open, 2025, 8: e2528056

37 Duggan M J, Gervase J, Schoenbaum A, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open, 2025, 8: e2460637

38 Lawrence K, Kuram V S, Levine D L, et al. Informed consent for ambient documentation using generative AI in ambulatory care. JAMA Netw Open, 2025, 8: e2522400

39 Longhurst C A, Singh K, Chopra A, et al. A call for artificial intelligence implementation science centers to evaluate clinical effectiveness. NEJM AI, 2024, 1: AIp2400223