SCIENCE CHINA
Information Sciences

• Supplementary File •

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, and 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 Women and Children Medical Care Center, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430010, China*
[3] *Lazaridis School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario, 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*

Appendix A Study oversight, reporting, and registration (Appendix 1)

This study was approved by the Institutional Review Board (IRB) of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (Approval No. [2025] S024). Written informed consent was obtained from all participants prior to enrollment. The study was conducted in accordance with the Declaration of Helsinki and applicable local regulations.

The study was reported following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for observational studies and the Chatbot Assessment Reporting Tool (CHART) considerations for studies involving conversational AI systems.

Appendix B System architecture (Appendix 2)

Shuwen employs a multi-agent orchestration framework comprising five specialized agents: (1) an identity and EHR reconciliation agent, (2) a medical history elicitation agent, (3) a pain management and perioperative counseling agent, (4) a risk assessment agent utilizing graph-enhanced retrieval-augmented generation (GraphRAG), and (5) a structured note generation agent. The system supports dual-model operation: a cloud-based GPT-4o model for complex reasoning and a locally deployed Qwen-2.5-14B model for on-premises processing of protected health information. The perioperative knowledge graph was constructed from standardized clinical documents, pharmacological databases, and procedure-specific protocols, cross-validated against the Chinese Medical Knowledge Graph (CMeKG).

Appendix B.1 Agent implementation and scheduling

The five agents are prompt-defined functional roles operating over a shared patient state, rather than five independently trained language models. A deterministic top-level scheduler advances through the predefined clinical stages, while responses within a stage can trigger additional role-specific questions. The scheduler supplies the active role with the accumulated dialogue state, relevant required fields, scope restrictions, and any retrieved evidence. This hybrid design preserves a reproducible clinical sequence while allowing patient-dependent follow-up.

Appendix B.2 Dialogue state and structured output

The shared state links the current interview stage, required and optional information slots, patient responses, unresolved items, retrieved evidence, and the draft structured note. Core slots cover identity and demographics, comorbidities, prior surgery and anesthesia, medications, allergies, lifestyle factors, airway-related information, fasting status, analgesia preferences, and risk-relevant symptoms. Information not provided by the patient remains explicitly unresolved rather than being inferred. Completion requires the mandatory interview stages to be traversed; unresolved or high-risk information is carried forward for anesthesiologist verification before sign-off.

* Corresponding author (email: dinghan@hust.edu.cn, huixu@tjh.tjmu.edu.cn, zht@mail.hust.edu.cn)
† Hong Tao, Yi-Fan Hu, and Hui Yu contributed equally to this work.

Appendix B.3 Knowledge graph and GraphRAG

The anesthesia-focused graph contained approximately 12,000 nodes, including about 11,000 entities, approximately 50,000 relations, about 80 relation types, and approximately 30–50 entity classes. Principal entity classes include diseases, symptoms, medications, treatments, medication-management advice, and perioperative risks. Candidate concepts and relationships extracted from perioperative forms, pharmacological guidance, and procedure-specific protocols were aligned with CMeKG terminology, checked for semantic consistency, and manually reviewed when ambiguous. Additional anesthesia knowledge was incorporated during deployment after clinical-expert review. Retrieval is activated during risk-assessment and medication-guidance stages. For each query, five candidate items are retrieved by combining vector retrieval with graph indexing; the corresponding graph triples are inserted into the active role prompt together with the accumulated patient state, allowing the generated explanation or recommendation to be linked to a bounded domain context.

Appendix B.4 Deployment, privacy, and auditability

During system development, alternative model configurations were explored and prompt templates were iterated; the reported clinical evaluation used a fixed local Qwen-2.5-14B checkpoint rather than switching model weights during the evaluation. Protected health information was processed on premises using Qwen-2.5-14B hosted on a server with four NVIDIA 4090 GPUs. External large models could be used only for deidentified text, such as questions about potential medication harms or questions outside the predefined interview workflow. Direct identifiers, including names, hospital numbers, identification numbers, telephone numbers, and addresses, were removed primarily by automated rules with limited manual review by study researchers. Operational audit records supported review of the active agent or interview stage, patient and generated response text, retrieved evidence or evidence identifiers, uncertainty or risk flags, retries and network or speech-recognition errors, and clinician takeover, handoff, modification, and final sign-off. The clinical workflow retained anesthesiologist review, targeted physical examination, correction of discrepancies, and final sign-off as mandatory steps.

Appendix B.5 Domain adaptation and failure handling

The local Qwen-2.5-14B model was adapted to perioperative dialogue using doctor–patient preoperative communication records. The adaptation data were split 9:1 into training and test subsets. LoRA fine-tuning used a 4-bit quantized base model with rank 16. Role prompts and shared state constrained each task at inference time. During implementation, the approximate end-to-end response delay was 5 s, comprising approximately 0.8–1.0 s for speech recognition, 3.0–3.2 s for model inference, and 0.8–1.0 s for speech synthesis. Safety controls included role and scope restrictions, one-question-at-a-time elicitation, explicit treatment of missing information, graph-scoped retrieval, predefined escalation conditions, and clinician confirmation. Patient preference for direct physician contact, network interruption, or unresolved safety-critical information triggered transition to the physician-led pathway. Potential implementation failures include speech-recognition error, network interruption, irrelevant retrieval, repeated questioning, and incomplete structured fields; these were managed through confirmation questions, bounded retries, preserved unresolved fields, and clinical handoff rather than autonomous resolution.

Appendix C Prompting strategy (Appendix 3)

The prompting framework follows four core principles: (1) Role Anchoring—each agent operates within a clearly defined clinical role with explicit scope boundaries; (2) Sequential Elicitation—structured questioning follows a clinically validated sequence to ensure comprehensive information gathering; (3) Clarity and Safety—all responses are calibrated for patient comprehension and include appropriate safety disclaimers; (4) Contextual Awareness—the system maintains awareness of previously gathered information to avoid redundant questioning and ensure clinical coherence.

The multi-agent prompting framework assigns each agent a specific functional role with illustrative prompts designed to guide clinically appropriate dialogue while maintaining patient safety.

Each prompt template included: (1) the agent's clinical role and permitted scope; (2) the current dialogue stage and accumulated patient context; (3) required information slots and instructions for handling missing or contradictory information; (4) retrieved graph evidence when applicable; (5) response-style constraints, including plain language and one question at a time; and (6) completion, escalation, and clinician-handoff conditions. These elements were intended to constrain the model to the active task while preserving a coherent longitudinal interview.

Appendix D Accuracy assessment (Appendix 4)

Appendix D.1 Reference standard

The anesthesiologist-verified final chart served as the reference standard for all accuracy assessments. Three senior anesthesiologists (each with >10 years of clinical experience) independently evaluated AI-generated summaries while blinded to one another's ratings. 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. Because the final chart was completed after review of the AI-generated draft within the clinical workflow, it was not a fully independent de novo reference standard; the accuracy endpoint is therefore interpreted as concordance with the clinician-verified final record, with potential reviewer familiarity and workflow-related bias acknowledged.

Appendix D.2 Consensus algorithm

A majority rule approach ($\geqslant$2 of 3 raters) was used to determine accuracy for each category and the overall document. A document was considered correct by a given rater only if all three categories (medications, diagnoses, and allergies) were rated as "correct."

Appendix D.3 Interrater agreement

Interrater agreement was quantified using Gwet's AC1 statistic, which provides a more stable estimate than Cohen's kappa in the presence of high agreement. Gwet's AC1 is defined as:

$$\text{AC1} = \frac{p_a - p_e}{1 - p_e}$$

where p_a denotes observed agreement and p_e denotes chance agreement.

Appendix D.4 Accuracy assessment rulebook

The assessment rulebook defines binary correctness criteria for three categories:

- **Medications**: All current medications (including dose, frequency, and route) are correctly documented without omissions or additions.
- **Diagnoses**: All active diagnoses relevant to perioperative management are correctly identified and documented.
- **Allergies**: All known allergies (including substance and reaction type) are correctly recorded without omissions or additions.

These three domains were prespecified because they are safety-relevant, routinely documented, and consistently comparable with the final clinical record. A category was marked incorrect for a clinically relevant omission, unsupported addition, or substantive mismatch in the required content. The composite did not measure all elements of preanesthetic assessment; prior anesthesia history, airway-related information, fasting status, and patient preferences remained part of the interview and clinician review but were not included in the primary accuracy endpoint.

Appendix E Patient and clinician surveys (Appendix 5)

Appendix E.1 Patient survey

The 14-item patient questionnaire covers five domains: overall satisfaction (3 items), interaction efficiency (3 items), information clarity (3 items), empathy (3 items), and perceived professional competence (2 items). All items are rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The survey was adapted from the Patient Satisfaction Questionnaire Short Form (PSQ-18) and tailored to perioperative AI interactions.

Appendix E.2 Clinician survey

The 16-item clinician questionnaire covers eight domains: overall satisfaction, accuracy, safety, communication quality, practicality, efficiency, accessibility, and preference for continued use. All items are rated on the same 5-point Likert scale.

Appendix E.3 Confirmatory factor analysis

CFA was conducted using WLSMV estimation on polychoric correlations. Model fit was assessed using CFI/TLI $\geqslant$0.95, RMSEA $\leqslant$0.08, and SRMR $\leqslant$0.08. Internal consistency was evaluated using Cronbach's α and composite reliability.

Appendix F Statistical analysis plan (Appendix 6)

The statistical analysis plan was prespecified prior to data analysis. Primary analyses included: (1) document-level accuracy with 95% Clopper-Pearson confidence intervals; (2) category-level accuracy stratified by ASA status; (3) interrater agreement using Gwet's AC1; (4) patient satisfaction comparisons using independent samples t-tests with Holm-adjusted P values; (5) clinician evaluation descriptive statistics; (6) Turing test indistinguishability rate. All tests were two-sided with $\alpha = .05$. Analyses were conducted using R version 4.3.2 (packages: irrCAC, psych, lavaan) and Python version 3.11 (packages: statsmodels, scipy).

Appendix G Quality assurance and monitoring (Appendix 7)

Quality assurance measures included: (1) standardized training for all study personnel; (2) pilot testing of the interview system with 20 patients before formal enrollment; (3) daily monitoring of system performance logs; (4) weekly review of adverse events or protocol deviations by the study team; (5) independent data monitoring by the institutional IRB. No serious adverse events related to the AI system were observed during the study period.

Appendix H Reproducible code and data availability (Appendix 8)

Reproducible analysis code is available upon reasonable request. Deidentified patient transcripts from the conversational AI interviews are not publicly available to protect patient privacy, in accordance with Tongji Hospital ethics and governance policies. Data access requests may be directed to the corresponding author (Prof. Hui Xu, huixu@tjh.tjmu.edu.cn).

References

1 Harms C, Kindler C H. The preoperative anaesthetic visit. Ther Umsch, 2009, 66: 503–508
2 Jindal P, Patil V, Pradhan R, et al. Update on preoperative evaluation and optimisation. Indian J Anaesth, 2023, 67: 39–47
3 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
4 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
5 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