OPEN ACCESS
Digital Medicine COMMENTARY
The Innovation Informatics

From language to world: Bridging LLMs and world models for intelligent surgery

Yang Liu,[1,2] Yuxuan Song,[1,2,*] Yiqing Du,[1] Caipeng Qin,[1] and Tao Xu[1,*]
[1]Department of Urology, Peking University People's Hospital, Beijing 100044, China
[2]These authors contributed equally
*Correspondence: yuxuan_song@bjmu.edu.cn (Y.S.); xutao@pkuph.edu.cn (T.X.)
Received: January 21, 2026; Accepted: April 17, 2026; Published Online: April 18, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100040

Citation: Liu Y., Song Y., Du Y., et al. (2026). From language to world: Bridging LLMs and world models for intelligent surgery. *The Innovation Informatics* **2**:100040.

Surgical artificial intelligence (AI) is moving in two directions that rarely meet. Large language models (LLMs) can read guidelines and clinical notes and produce reasonable-sounding advice, but they cannot show that a step is safe for a specific patient at a specific moment. World models and surgical digital twins can simulate tissue and tool behavior, but they often lack a clear way to express clinical intent and to explain why one option is chosen over another. Treating these systems separately leads to advice without physical checks, or simulation without clinical meaning. We argue for a cognitive-physical loop. LLMs should translate goals into simple task steps and constraints, and patient-specific models should test these proposals and return safety limits that can guide action in the operating room.

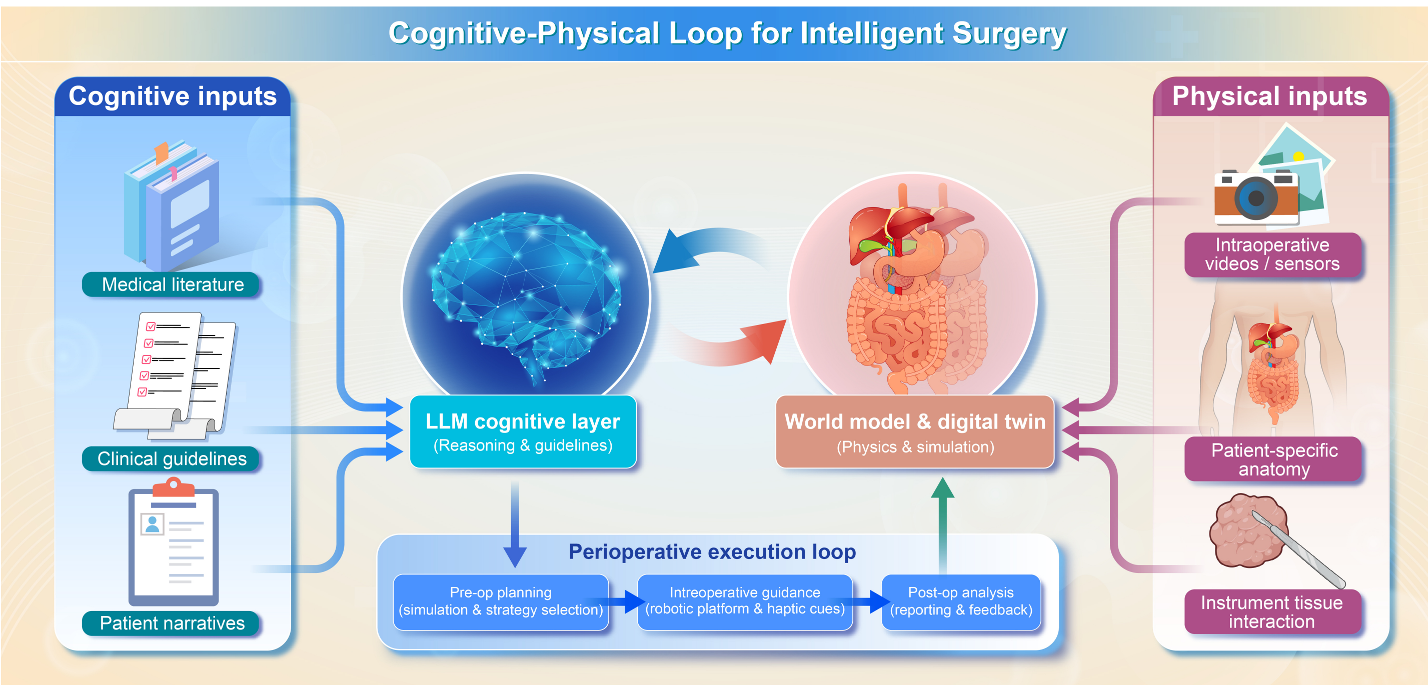

Figure 1. The cognitive-physical loop for intelligent surgery.

FROM DIGITAL TO INTELLIGENT SURGERY

Over the past three decades, surgery has become increasingly digital. Robots, image guidance, and intraoperative imaging have improved precision and made minimally invasive procedures easier. Electronic health records have also made clinical data easier to access. Yet most AI used in the operating room is still narrow and built for one task, such as image analysis or phase recognition.[1]

Robotic systems are also becoming harder to assess. The updated IDEAL framework for AI-enabled surgical systems argues that evaluation should cover not only technical accuracy but also workflow, human factors, cost, and long-term safety.[1] This shift points to a new goal. We need systems that can support decisions and actions in the real operative setting.

We see two building blocks for this step. LLMs can work with clinical language and help turn evidence and notes into structured goals and constraints.[2] World models and digital twins can represent anatomy and tool interactions over time and can test what might happen under different actions in patient specific settings.[3] Our main point is simple. Intelligent surgery is more likely to come from linking these two parts than from improving either part alone.

LARGE LANGUAGE MODELS AS THE COGNITIVE LAYER

In surgery, LLMs can reduce paperwork and support training, but their bigger role is to turn clinical text into information that a team can act on.[2] Before surgery, an LLM can summarize comorbidities and prior reports, then compare them with guidelines and trial evidence to surface key options for a given patient. During surgery, it can track the plan and the current context and flag important deviations. After surgery, it can help draft operative notes, discharge letters, and patient instructions, which can reduce documentation time and improve consistency.[2]

Domain adaptation matters. SurgeryLLM links a general LLM to a curated set of surgical guidelines and patient data.[4] In simulated cardiovascular cases it did better than a base model at finding missing workups, flagging abnormal results, and suggesting care that matched guidelines. This supports a practical rule for surgical use. LLMs should be tied to clear sources that others can check, not used as standalone reasoning engines.[2]

Even with retrieval, limits remain. LLMs can produce answers that sound right but are wrong, especially for rare cases or when the needed evidence is not in the source set.[2] In an intelligent surgery system, LLM output should not stay as free text. It should be converted into structured steps, constraints, or alerts that can be reviewed and logged. The model should be allowed to say when evidence is weak. When uncertainty is high, the system should fall back to full human control.

WORLD MODELS AND DIGITAL TWINS AS THE PHYSICAL LAYER

LLMs work well with text, but surgery happens in a physical setting. World models and digital twins aim to capture that setting. In digital twin-assisted surgery, each patient is represented by a virtual model that combines imaging with anatomical and physiological data.[3] This model can support planning before surgery, guidance during surgery, and review after surgery. It can help teams compare resection margins, reconstruction options, and instrument paths in simulation, and it can help predict how a specific anatomy and tissue behavior may respond to an action.[3,5] In this view, a surgical digital twin is a patient-specific world model that links actions to changes such as blood flow, tissue deformation, and bleeding.

World models also support training for robotic autonomy. With realistic simulators, robots can practice basic subtasks such as tissue retraction and exposure. Policies learned in simulation can sometimes be transferred to animal surgery with limited adjustment, which shows how simulation can be used for safer exploration and skill learning.

A major challenge is the bridge from clinical advice to robot control. LLMs should not output joint angles or force curves. Instead, they should work with a small set of validated surgical skills and task steps. World models and digital twins can then map these steps to safe trajectories that respect patient specific anatomy and safety limits. Building and maintaining patient-specific twins is not easy. It needs high-quality multimodal data, timely updates during surgery, and careful validation. For this reason, it makes sense to start with use cases where the added prediction clearly outweighs the added complexity.[3,5]

THE COGNITIVE-PHYSICAL LOOP: A UNIFIED ARCHITECTURE

We describe intelligent surgery as a cognitive-physical loop that links LLMs, world models, and the surgical team. Patient-specific models simulate how tissue and instruments may behave under a proposed action and they estimate risk. The LLM layer turns these predictions, together with guidelines and patient data, into clear options and short explanations. The execution layer includes the robot and navigation tools that act under surgeon control. The overall architecture of this cognitive-physical loop is illustrated in Figure 1.

The system integrates two core layers. The cognitive layer (left) utilizes Large Language Models (LLMs) to process medical text, guidelines, and patient narratives, providing reasoning and strategic planning. The physical layer (right) employs world models and digital twins to simulate patient-specific anatomy and instrument-tissue interactions based on sensor data. These two layers interact in a closed loop: the cognitive layer proposes tasks, and the physical layer validates them within a simulated physical environment. This architecture supports the full perioperative execution loop, enabling pre-operative planning, real-time intraoperative guidance, and post-operative analysis.

This loop can support planning and intraoperative support. In planning, the LLM summarizes key comorbidities and constraints from the record and maps them to guideline-based options. The digital twin then tests candidate strategies such as different resection planes or reconstruction plans and reports expected margins, organ preservation, and complication risk. The LLM uses these outputs to present tradeoffs for the team in plain language. During surgery, the physical model updates with video and sensor data and can estimate tool distance to critical structures. It can also predict what may happen if a maneuver is attempted. When risk rises, the system can trigger short alerts and request overlays that highlight vulnerable structures and safer paths. For narrow subtasks, controllers trained in simulation may assist under surgeon approval, and they should hand over immediately if conditions change.

In this loop, the LLM is not only a narrator. It can propose options within clear limits by asking the physical model to test alternative task steps and then comparing their predicted risks. Safety rules are essential. The team should be able to review each update, and the system should show what comes from model prediction and what comes from guidelines.

EVALUATION, GOVERNANCE AND EDUCATION

Intelligent surgical systems that can both suggest and support action need evaluation that goes beyond classic device trials. The updated IDEAL framework for surgical robotics calls for staged assessment from early concepts to long-term monitoring. It also asks teams to report clinical outcomes, cost, and human factors.[1] For systems that link LLMs with world models, we also need to test how errors spread across layers. A weak physical model can give wrong predictions, and an LLM can present these predictions with too much confidence. We should measure not only each part on its own but also how the full human-AI loop behaves in real workflows.

Governance matters as much as performance. Policy papers raise questions about responsibility when AI contributes to harm. They also ask how to document model behavior and how to include AI tools in informed consent. Coupled systems make these issues sharper because advice and action sit close together. A practical step is to design for auditability. Systems should keep structured logs of key inputs, model settings, simulation results, and what was shown to the team. Logs should also record clear human sign-off points so that responsibility stays with named clinicians and institutions.

Education will shape safe adoption. Surgeons will need basic AI literacy. They should know what these models can do, what they cannot do, and when to doubt them. Digital twins can also help training by offering repeatable scenarios and feedback. If these tools are used in training, they should support clinical judgment rather than replace it.

REFERENCES

1. Marcus H.J., Ramirez P.T., Khan D.Z., et al. (2024). The IDEAL framework for surgical robotics: Development, comparative evaluation and long-term monitoring. *Nat. Med.* **30**:61–75. DOI:10.1038/s41591-023-02732-7
2. Thirunavukarasu A.J., Ting D.S.J., Elangovan K., et al. (2023). Large language models in medicine. *Nat. Med.* **29**:1930–1940. DOI:10.1038/s41591-023-02448-8
3. Asciak L., Kyeremeh J., Luo X., et al. (2025). Digital twin assisted surgery, concept, opportunities, and challenges. *NPJ Digit. Med.* **8**:32. DOI:10.1038/s41746-024-01413-0
4. Ong C.S., Obey N.T., Zheng Y., et al. (2024). SurgeryLLM: A retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *NPJ Digit. Med.* **7**:364. DOI:10.1038/s41746-024-01391-3
5. Mekki Y.M., Luijten G., Hagert E., et al. (2025). Digital twins for the era of personalized surgery. *NPJ Digit. Med.* **8**:283. DOI:10.1038/s41746-025-01575-5

FUNDING AND ACKNOWLEDGMENTS

This study was supported by National Key Research and Development Program of China (2023YFC2507000), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0525700), Innovation Fund for Outstanding Doctoral Candidates of Peking University Health Science Center (BMU2024BSS001), National Natural Science Foundation of China (82471866, 82271877, 82472912, 82371840), Natural Science Foundation of Beijing, China (7242150, 7262134, 7264343, QY25150), Beijing Municipal Science & Technology Commission (Z221100007422097), Capital's Funds for Health Improvement and Research of China (2022-4-4087), Peking University People's Hospital Scientific Research Development Funds (RDGS2022-02, RDX2024-01, RDEB2025-02, RDX2025-01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: Y.L., Y.S. Resources: Y.D. Visualization: Y.D., C.Q. Writing - Original Draft: Y.L. Writing - Review & Editing: Y.S. Supervision: T.X. Project administration: T.X. Funding acquisition: T.X. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.