Beyond Benchmarks: A Capability-Based Maturity Model for Systematic AI Integration in Hospitals

Rui Li[1,*] , Xiaofen Wu[2,*] , Mingqian Liu[1], Xiaoxia Song[1], Xiangjun Xu[1], Jiacheng Qiao[1], Boqin Zhuang[1], Xu Chen[1,†]
[1]Winning Health AI Research, Winning Health Technology Group Co., Ltd
[2]Information Department, Tongji University Affiliated Tongji Hospital

Abstract

Current Large Language Models (LLMs) demonstrate exceptional performance on medical benchmarks. However, models that excel in standardized tests focused on medical knowledge recall are not necessarily effective in real-world healthcare scenarios. This disparity between academic performance and clinical effectiveness stems from existing evaluations focusing overly on knowledge retrieval and QA, while neglecting high-load executive tasks in real clinical workflows. The effective execution of such tasks depends not only on model reasoning but also on the overall digital maturity of the healthcare institution. To address this, we propose a "Capability-Based Hospital AI Maturity Model" framework. This framework establishes a layered maturity system based on capabilities. By categorizing hospital AI capabilities into distinct maturity levels, it provides a clear, stepwise evolutionary path for hospitals, guiding them from foundational infrastructure construction to ubiquitous intelligence. Guided by this framework, we constructed ten representative real-world clinical scenarios as a reference test set and compared the performance of multiple models across benchmarks and real-world scenarios. Preliminary results suggest that, compared to relying solely on academic benchmark scores, this maturity assessment mode—which integrates system governance and scenario constraints—may provide a more valuable basis for AI adoption in medical institutions.

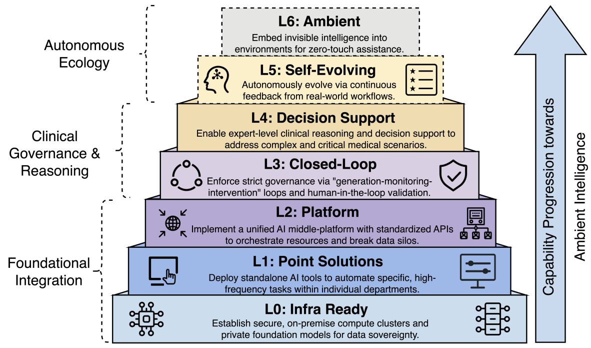

Figure 1: The Capability-Based Hospital AI Maturity Model.

1 Introduction

Medical Artificial Intelligence is currently at a critical transition stage from technical verification to deep clinical integration (Aravazhi et al., 2025; Topol, 2019). Although general LLMs (OpenAI et al., 2024; Touvron et al., 2023) and medical-specific models continue to break records on influential public benchmarks, we must face a reality in actual deployment: models that perform excellently in standardized test environments often fail to translate directly into expected clinical value when intervening in complex real-world hospital workflows (Bedi et al., 2025).

This misalignment essentially stems from the singularity of evaluation dimensions. A model skilled only in knowledge recall and exam-oriented QA, if lacking adaptability to hospital private data distributions or unable to strictly follow clinical specific quality control and compliance instructions, cannot undertake real medical work. In other words, clinical scenarios require not just the model's reasoning ability, but the system capability of the model to deeply integrate with existing hospital processes and governance systems. However, building such systematic capabilities is highly complex. Currently, due to the lack of unified evaluation standards and construction paths, many medical institutions find it difficult to precisely locate their governance shortcomings when advancing AI transformation.

Addressing this core pain point, this paper primarily contributes the following three aspects:

First, we propose a Construction Guidance Framework. We introduce the "Capability-Based

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20883–20895

Hospital AI Maturity Model," offering a stepwise evolutionary path from infrastructure to ambient intelligence. This framework moves beyond single-dimensional performance metrics, providing a systematic structure for AI adoption (HIMSS, 2024; ISO, 2023).

Second, we constructed Real Verification Scenarios. We designed 10 representative clinical tasks, ranging from administrative execution to decision support. These serve as reusable templates for medical institutions to build localized evaluation benchmarks aligned with their specific workflows (Zhang et al., 2022).

Third, we conducted an Empirical Evaluation of Discrepancies. Our experiments reveal that high academic scores do not guarantee clinical effectiveness. We highlight how models often fail in real scenarios due to instruction compliance or reasoning deficits, emphasizing the critical role of system governance architectures.

2 Related Work

2.1 Medical AI Benchmarks

As LLM application in the medical field deepens, high-quality evaluation benchmarks have driven quantitative research on clinical capabilities (Vaswani et al., 2023; Alsentzer et al., 2019). These works can be roughly divided into two evolutionary stages:

Phase 1: Static Reasoning Based on Medical Knowledge. Early benchmarks focused on measuring the model's memory of standardized medical knowledge and single-step reasoning. MedQA (Jin et al., 2021) derives directly from physician licensing exams in the US and Mainland China/Taiwan, requiring models to choose from 4-5 options, assessing professional access levels. MedMCQA (Pal et al., 2022) collects over 190,000 questions from Indian medical entrance exams, covering 21 disciplines. PubMedQA (Jin et al., 2019) adopts a literature-based QA format, testing logical judgment based on evidence-based medicine.

Phase 2: Comprehensive Ability and Interactive Evaluation. To address complex clinical realities, works have begun building comprehensive benchmarks involving multi-modality, multi-turn, and specific contexts. MedXpertQA (Zuo et al., 2025) positions itself at "expert-level" difficulty, introducing complex case-based board questions. MedBench (Liu et al., 2024) and HealthBench (Arora et al., 2025) focus on large-scale benchmarks for specific language contexts and interactive scenarios involving empathy and safety boundaries.

2.2 Hospital Informatics and Maturity Assessment

Existing assessment systems focus on infrastructure and digitization. HIMSS EMRAM & INFRAM (HIMSS, 2024) are global gold standards guiding hospitals from paperless records to interoperability and evaluating underlying hardware readiness. ISO/IEC 42001 (ISO, 2023) provides a normative framework for AI risk management.

3 Hospital AI Maturity Model Framework

To systematically deconstruct and guide the integration path of medical AI, we propose a seven-level capability-based maturity model framework (see Figure 1). This framework abstracts complex system evolution into three continuous stages.

3.1 Phase 1: Foundation and Platformization

The core task is breaking data silos and establishing standardized computing infrastructure.

L0: Infra Ready. Definition: The hospital possesses basic computing resources and has established foundation model capabilities through localized deployment or dedicated commercial APIs. Core Capability: Secure operation of models within the hospital intranet or dedicated private networks, ensuring data sovereignty. Value: Solves the "cold start" problem, guaranteeing basic privacy and compute supply.

L1: Point Solutions. Definition: AI serves specific department business scenarios as independent tools or workstations. Core Capability: Automated assistance for high-frequency, mechanical single tasks. Value: Quickly addresses specific pain points and improves efficiency without complex integration.

L2: Platform. Definition: Establishing a hospital-wide AI middle platform and API gateway. Core Capability: Unified scheduling of model resources, version management, and standard API encapsulation. Value: Breaks data chimneys, avoids repetitive procurement, reduces operations costs, and improves resource utilization.

3.2 Phase 2: Governance and Intelligence

This phase focuses on building a trustworthy environment, introducing strict quality control (QC) and high-order reasoning.

L3: Closed-Loop Governance. Definition: Introducing safety guardrails and deterministic rule engines to build a "Generation-Monitoring-Intervention" closed-loop QC system. Core Capability: Real-time interception and correction of model inputs/outputs. Value: Ensures AI behavior remains within hospital management regulations, reducing medical risk (Ouyang et al., 2022).

L4: Decision Support. Definition: Comprehensive analytical capability for complex clinical situations, bridging the gap to expert-level clinical thinking. Core Capability: Utilizing long-context and Chain-of-Thought (CoT) (Lewis et al., 2021) to combine surface information with authoritative literature/regulations for comprehensive judgment. Value: Compensates for cognitive limitations in processing massive information, assisting in high-risk critical decisions.

3.3 Phase 3: Symbiosis and Future

L5: Self-Evolving System. Definition: Establishing a "Data Flywheel" with continuous RLHF. Value: Performance continuously climbs via usage data accumulation.

L6: Ambient Intelligence. Definition: Invisible AI integration into physical space and digital flows (Zero-touch). Value: Reshapes hospital operations; technology recedes into the background as a ubiquitous safety foundation.

4 Experimental Design and Scenarios

To verify the framework, we constructed a test benchmark containing ten typical clinical tasks (L0-L4). Detailed definitions are in the Appendix. The scenarios are: **L0:** Standardized Medical Licensing Exam (CNMLE) and Clinical Terminology Completion (CTC). **L1:** Clinical Form Filling (CFF) and Radiology Report Generation (RRG). **L2:** Time Logic Validation (TLV) and Ultrasound-Radiology Consistency (UDC). **L3:** Rule-Based Record QC (RBQ) and Document Error Correction (DEC). **L4:** TNM Staging Decision (TNM) and MDT Suggestion Generation (MDT).

4.1 Public Benchmark Datasets

To build a complete evaluation reference frame, we selected MedQA (USMLE), MedMCQA, PubMedQA, and MedXpertQA (text) as academic baselines. These datasets represent standardized evaluations of knowledge recall and reasoning, serving as a control group to contrast model performance.

4.2 Evaluation Metrics

To adapt to the task differences, we adopted a classified evaluation system. Type A (Discrimination) uses Accuracy. Type B (Structured reasoning) uses Logical Accuracy. Type C (Open-ended generation) uses LLM as Judge. Specific details are provided in the Appendix.

5 Experiments

5.1 Experimental Setup

Given the high sensitivity of medical data and strict hospital requirements for data sovereignty, this experiment excludes all closed-source commercial API services. We selected open-source models capable of private deployment as evaluation objects, covering both general LLMs and medical-specific fine-tuned models. To ensure fairness, we used identical system prompts and default inference parameters for all models across the same tasks. To ensure robustness, all reported results are averaged over 5 independent iterations.

5.2 Performance Analysis

The multidimensional evaluation (Figure 2) and results in Table 2 reveal that clinical capability is not a linear function of benchmark scores. Instead, we observe three structural characteristics:

Asymmetric Capability Distribution. As visualized in Figure 2, model capabilities are heavily skewed towards the upper-right axes (L0: Infra Ready, L1: Point Solutions). While 7B-class models achieve near-saturation in standardized exams and simple extraction, their performance significantly degrades in the lower-left sectors (L3: Closed-Loop, L4: Predictive). This asymmetry confirms that while medical knowledge retrieval has become a commoditized capability, high-order clinical reasoning and strict constraint adherence remain emergent abilities that are strictly dependent on model scale.

Trade-offs in Domain Specialization. Our experiments indicate that medical-specific fine-tuning

Table 1: Model performance on public academic benchmarks.

Model	MedQA	MedMCQA	PubMedQA	MedXpertQA
gemma-3-4b-it (Team et al., 2025a)	49.00	42.82	47.40	11.29
Qwen3-4B-Instruct-2507 (Yang et al., 2025)	73.73	60.97	76.48	16.78
medgemma-4b-it* (Sellergren et al., 2025)	62.30	53.80	70.46	14.16
gpt-oss-20b (OpenAI et al., 2025)	84.90	67.14	77.04	25.95
gemma-3-27b-it (Team et al., 2025a)	74.37	63.18	42.28	15.42
medgemma-27b-text-it* (Sellergren et al., 2025)	87.56	73.18	72.12	26.59
Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025)	86.11	71.33	77.66	23.39
Baichuan-M2-32B* (Team et al., 2025b)	89.29	71.66	70.74	27.21
GLM-4-32B-0414 (GLM et al., 2024)	81.44	66.93	54.60	20.41
Seed-OSS-36B-Instruct (Team, 2025)	88.31	72.93	72.24	28.34
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	84.15	73.63	**79.48**	24.31
gpt-oss-120b (OpenAI et al., 2025)	91.81	74.67	76.88	**35.18**
Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025)	**91.96**	**78.17**	76.60	33.14

Table 2: Model performance across L0-L4 clinical maturity scenarios. (* denotes medical-specific models).

Model	L0		L1		L2		L3		L4	
	CNMLE	CTC	CFF	RRG	TLV	UDC	RBQ	DEC	TNM	MDT
gemma-3-4b-it	43.18	18.16	65.77	63.31	53.80	48.80	64.50	2.93	3.14	45.15
Qwen3-4B-Instruct-2507	84.83	44.88	84.94	65.08	68.20	38.80	63.40	42.87	20.00	80.39
medgemma-4b-it*	54.42	21.84	68.46	59.11	33.40	55.60	64.55	10.22	1.82	44.75
gpt-oss-20b	80.12	35.25	93.46	65.80	72.40	60.40	70.12	29.51	21.49	64.31
gemma-3-27b-it	75.09	34.71	90.38	71.04	37.60	50.60	59.80	12.13	5.46	69.56
medgemma-27b-text-it*	81.22	36.73	90.90	66.65	57.00	51.60	61.45	12.56	4.96	74.00
Qwen3-30B-A3B-Instruct-2507	90.06	58.34	92.05	71.78	61.80	40.60	65.67	32.13	23.64	80.11
Baichuan-M2-32B*	90.76	49.34	90.58	74.96	46.20	51.60	**72.92**	55.71	22.31	80.14
GLM-4-32B-0414	86.40	50.53	92.24	75.66	53.20	51.20	60.95	46.30	18.18	75.49
Seed-OSS-36B-Instruct	93.01	**67.38**	82.12	**78.47**	71.20	52.80	63.67	**60.83**	26.12	81.93
Llama-3.3-70B-Instruct	84.08	33.85	90.45	68.66	63.00	45.60	66.25	48.89	15.21	66.25
gpt-oss-120b	84.40	49.58	92.05	65.33	**76.20**	55.00	70.72	28.52	22.31	79.97
Qwen3-235B-A22B-Instruct-2507	**93.20**	66.74	**94.23**	75.17	59.00	**67.00**	62.33	36.98	**26.61**	**84.91**

often leads to performance regression in L3 Governance tasks. Specialized models frequently exhibit a retraction along the L3 axis compared to their base versions. This suggests that aggressive tuning on medical corpora may overfit to domain content, compromising the model's general robustness in following complex, negative constraints required for administrative quality control.

Divergence between Benchmarks and Reality. High academic scores do not guarantee real-world effectiveness. Models like Llama-3.3-70B excel in MedQA yet show limited coverage in L4 scenarios. Consequently, the total area covered in the radar chart serves as a more reliable proxy for the systematic maturity of a model than single-metric leaderboards.

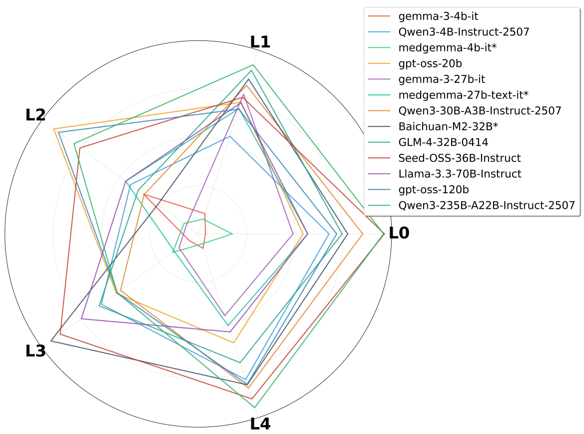

Figure 2: Radar Chart of Capability Envelopes (L0-L4).

6 Conclusion

This paper proposes the Capability-Based Hospital AI Maturity Model to guide the systematic construction of AI in medical institutions. Through empirical verification across ten clinical scenarios (L0-L4), we demonstrate that high benchmark scores often mask significant deficiencies in real hospital workflows.

Our analysis identifies a substantial disparity in model capabilities: while current systems excel in knowledge retrieval (L0), they face significant bottlenecks in complex instruction compliance (L3) and decision support (L4). The proposed framework provides a stepwise path for medical institutions to assess AI readiness, shifting focus from parameter competition to the construction of balanced, governable clinical intelligence systems.

Limitations

Although this study preliminarily verifies the effectiveness of the framework, future work still needs to be deepened in the following three dimensions. High-Order Stages: This study covers L0-L4 offline verification but has not touched upon L5 and L6, which require real-time feedback and physical perception. Future work will explore online (On-policy) evaluation. Metric Adaptation: Unified latency/throughput metrics were not set due to hospital heterogeneity; these should be customized during engineering implementation. Expert Review: We provided reference scenarios without passing scores. Final maturity certification should involve expert review mechanisms.

Ethical Considerations

All clinical data were sourced from the internal EHR of our collaborating hospital. All data underwent rigorous de-identification. Processing occurred exclusively within a secure, on-premise private environment with no external transmission. The study received IRB approval. No direct human subject experimentation was involved.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

P. S. Aravazhi, P. Gunasekaran, N. Z. Y. Benjamin, A. Thai, K. K. Chandrasekar, N. D. Kolanu, P. Prajjwal, Y. Tekuru, L. V. Brito, and P. Inban. 2025. The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. *Disease-a-Month*, 71(6):101882.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*, 333(4):319–328.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

HIMSS. 2024. Electronic medical record adoption model (emram).

ISO. 2023. Iso/iec 42001:2023 information technology — artificial intelligence — management system.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Preprint*, arXiv:2407.10990.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

ByteDance Seed Team. 2025. Seed-oss open-source models. `https://github.com/ByteDance-Seed/seed-oss`.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025b. Baichuan-m2: Scaling medical capability with large verifier system. *Preprint*, arXiv:2509.02208.

Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, and 4 others. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. *Preprint*, arXiv:2106.08087.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Dataset Statistics and Scenario Definitions

We constructed ten clinical scenarios across maturity levels L0 to L4. Below are the detailed definitions and dataset sizes (**N**) for each scenario.

L0: Infra Ready (Knowledge Foundation)

- **CNMLE (N=509):** *Standardized Medical Licensing Exam.* Questions sourced from the Chinese National Medical Licensing Examination to test fundamental medical knowledge recall.

- **CTC (N=149):** *Clinical Terminology Completion.* A task assessing the mastery of fundamental anatomical knowledge. It requires the model to generate precise professional terminology based on standard textbook descriptions of anatomical structures..

L1: Point Solutions (Automation)

- **CFF (N=312):** *Clinical Form Filling.* Extract structured key-value pairs (e.g., patient age, symptoms, diagnosis) from unstructured admission notes into a strict JSON format.

- **RRG (N=300):** *Radiology Report Generation.* Generate a standard radiology finding report based on provided imaging features and diagnostic conclusions.

L2: Platform (Interoperability)

- **TLV (N=100):** *Time Logic Validation.* Identify logical contradictions in medical records regarding timestamps (e.g., discharge time cannot be earlier than admission time).

- **UDC (N=100):** *Ultrasound-Radiology Consistency.* Verify if the semantic conclusion of an Ultrasound report is consistent with the Radiologist's summary.

L3: Closed-Loop Governance (Safety)

- **RBQ (N=100):** *Rule-Based Quality Control.* Check medical records against explicit hospital rules.

- **DEC (N=108):** *Document Error Correction.* Identify and correct typos or semantic errors in clinical documentation while maintaining the original medical intent.

L4: Decision Support (Complex Reasoning)

- **TNM (N=121):** *TNM Staging Decision.* Determine the T, N, and M stages of cancer patients based on complex, multi-page pathology and surgical reports.

- **MDT (N=50):** *MDT Suggestion Generation.* Generate comprehensive multidisciplinary treatment plans for complex oncology cases, integrating guidelines and patient history.

B Representative Dataset Examples

The following figures present one representative example from each of the ten evaluation datasets (L0–L4), translated from the original Chinese. Long inputs are truncated with `[...]` for brevity.

C Evaluation Metrics and Prompts

To adapt to the diverse nature of tasks, we adopted a classified evaluation system:

Type A: Discrimination and Selection. Applied to tasks with a single unique standard answer (**CNMLE**). We utilize **Accuracy** as the core metric.

Type B: Structured Reasoning and Verification. Applied to tasks requiring structured decision outputs or specific format constraints (**CTC, CFF, TLV, UDC, RBQ, DEC**). We employ **Logical Accuracy**. A sample is considered correct only if the output strictly follows the required format (e.g., JSON or specific terminology) and the key decision fields match the ground truth.

Type C: Open-Ended Generation. For open-ended tasks involving complex reasoning and long-form generation (**RRG, TNM, MDT**), we adopt a **Model-Based Evaluation (LLM-as-a-Judge)** framework. We use a calibrated expert scoring model as the judge. The specific prompt template is provided below.

LLM-as-a-Judge Prompt Template

(a) CNMLE — Standardized Medical Licensing Exam

```
[Question]
A 4-year-old boy with blurred night vision for 6 months. Usually a picky
eater on a vegetarian diet, prone to respiratory infections and diarrhea.
PE: Bitot's spots visible on the lateral cornea of both eyes.
The most likely deficient vitamin is:

A. Vitamin A    B. Vitamin B1    C. Vitamin B2
D. Vitamin D    E. Vitamin C

[Answer] A
```

(b) CTC — Clinical Terminology Completion

```
[Question]
The superficial fascia contains ______, ______,
superficial lymphatic vessels, and cutaneous nerves.

[Answer] Superficial arteries, Superficial veins
```

Figure 3: Representative examples for **L0: Infra Ready** scenarios.

(a) CFF — Clinical Form Filling

```
[Task]
Analyze whether the indicator keyword appears in the patient data.
Follow the decision logic below:
  if (input data is empty or invalid) -> result: 0
  else if (surgery involves cranial/thoracic/abdominal regions)
    if (anesthesia method is general anesthesia) -> result: 1
    else -> result: 0
  else -> result: 0

[Indicator] Cranial-Thoracic-Abdominal General Anesthesia Surgery

[Patient Data]
Multi-catheter coronary angiography + PCI if necessary;
Multi-catheter coronary angiography; -; Local anesthesia

[Answer]
[{"indicator_name":"YYHDXFY-lxfbqmss","indicator_result":0}]
```

(b) RRG — Radiology Report Generation

```
[Procedure] ERCP

[Findings]
After successful general anesthesia, scope was advanced. Upper GI tract
unremarkable but significant food residue in stomach and duodenum.
Standard papilla with granular orifice; small periampullary diverticulum.
Guidewire entered pancreatic duct; re-cannulation into bile duct.
10ml iohexol injected; cholangiogram showed mild CBD dilation with
suspected oval filling defect. Sphincterotomy performed, balloon
dilation of papillary orifice to 1.0cm. Basket extraction of an
intact 1.2cm stone with bile sludge drainage. Submucosal hemostasis
performed. Nasobiliary drain and 5F x 6cm pancreatic stent placed.

[Task] Generate the examination conclusion.

[Answer]
ERCP + EST + balloon dilation + basket lithotomy + submucosal
hemostasis + ENBD + ERPD: Common bile duct stone; periampullary
diverticulum; food residue retention.
```

Figure 4: Representative examples for **L1: Point Solutions** scenarios.

(a) TLV — Time Logic Validation

```
[Task]
Evaluate whether the onset time stated in the chief complaint is
consistent with the admission date and the disease timeline described
in the present illness history.

[Admission Time] 2025-06-24 09:01:00
[Chief Complaint] Anti-tuberculosis treatment for 5 months.

[Present Illness (excerpt)]
Patient reports coughing with sputum 16 months ago. Chest CT showed
bilateral pulmonary inflammation [...] On 2024-12-29, patient visited
Xinhua Hospital: temperature 38.6C, CRP 101, chest CT showing bilateral
pneumonia with left pneumothorax and pleural effusion. Suspected
pulmonary TB; anti-tuberculosis therapy initiated with isoniazid,
rifampicin, levofloxacin IV, and ethambutol PO [...]

[Answer] false
(Anti-TB treatment started ~2024-12-29, approximately 6 months before
admission, not 5 months as stated in the chief complaint.)
```

(b) UDC — Ultrasound-Radiology Consistency

```
[Radiology Report Conclusion]
1. Multiple hepatic cysts.
2. Small left renal cyst.

[Ultrasound Report Conclusion]
Mild fatty liver; hepatic cyst. Please correlate clinically.

[Task] Determine if the two reports are semantically consistent.

[Answer] true
```

Figure 5: Representative examples for **L2: Platform** scenarios.

(a) RBQ — Rule-Based Quality Control

```
[Task]
Evaluate the patient's medical record against the hospital's
standardized quality control scoring criteria (excerpt below).

[Scoring Criteria (excerpt)]
| Item           | Requirement                    | Deduction       |
|----------------|--------------------------------|-----------------|
| Chief Complaint| Concise, <=20 chars, must      | -2 if unfocused |
|                | lead to primary diagnosis      | -1 if uses dx   |
| Present Illness| Consistent with chief complaint| -2 if unrelated |
|                | Onset time, symptoms, signs    | -1/item missing |
| Physical Exam  | Complete and consistent        | -1/item missing |
| Diagnosis      | Reasonable, properly ordered   | -2 if absent    |
[...]

[Patient Record (excerpt)]
Male, 67. Chief Complaint: Right axillary mass with pain for 1 week.
Present Illness: Patient discovered right axillary mass with pain
one week ago. Mass gradually enlarged with worsening pain. Ultrasound:
enlarged lymph node likely (inflammatory). Admitted as "right axillary
lymphadenitis." No fever, no night sweats. [...]

[Answer] Quality control scoring with itemized deductions.
```

(b) DEC — Document Error Correction

```
[Task]
You are a senior radiology QC expert. Identify and correct errors
in the report (e.g., typos, date errors, value/unit errors,
content duplication, inconsistencies). Preserve original intent.

[Report]
Findings: Normal skull shape. Patchy low-density shadow in/and the
left frontoparietal lobe, clear boundary. Ventricular system not
enlarged. Midline structures show no significant deviation. [...]
Left lower lobe lateral basal segment: solid nodule (Img192/290),
approximately 3.0mm x 2.7mm. [...]
Conclusion: 1. Left frontoparietal lobe encephalomalacia.
2. Left lower lobe solid nodule; follow-up in 6-12 months.

[Answer]
{"F": [{"error_text": "Findings: **in/and the left frontoparietal
  lobe patchy low-density shadow**",
  "correction": "Findings: a patchy low-density shadow is seen in
  the left frontoparietal lobe",
  "reason": "Terminology error: the Chinese character for 'and' was
  used instead of 'seen/present', altering the intended meaning."}]}
```

Figure 6: Representative examples for **L3: Closed-Loop Governance** scenarios.

(a) TNM — TNM Staging Decision

```
[Task]
You are a senior oncologic imaging expert. Based on the imaging report
below, perform TNM staging per AJCC 8th Edition.

[Report] Upper Abdominal CT with Contrast
Findings: Multiple ring-enhancing lesions in the liver parenchyma,
largest ~29mm, with blurred margins. No intra/extrahepatic bile duct
dilation. A 58x27mm hypodense mass in the pancreatic tail with mild
heterogeneous enhancement; tumor partially invading the spleen, with
indistinct boundary from adjacent gastric wall. Splenic artery and
vein involvement. Wedge-shaped hypodensity in spleen. Portal vein
and branches dilated. No enlarged abdominal lymph nodes. No ascites.

[Answer]
{"T_classification": "T4",
 "N_classification": "N0",
 "M_classification": "M1",
 "Stage_classification": "IV"}
```

(b) MDT — Multidisciplinary Team Suggestion Generation

```
[Task]
You are a member of the MDT expert panel for pregnancy complicated
by cervical cancer. Provide treatment recommendations from four
perspectives: Gynecologic Oncology, Chemotherapy, Obstetrics,
and Neonatology.

[Patient Summary]
- Diagnosis: Cervical cancer, FIGO Stage IB2
- Gestational age: 26 weeks
- Pathology: Squamous cell carcinoma (ordinary type)
- Patient's wish: Continue pregnancy

[Clinical Guidelines Provided]
- NACT regimen: Cisplatin + Paclitaxel or Carboplatin + Paclitaxel
  q3w; last cycle >= 3 weeks before delivery; not beyond 34 weeks.
- For IB2, pregnancy <22 weeks: Option (a) Laparoscopic pelvic
  lymphadenectomy first; Option (b) NACT then lymphadenectomy.
- For IB2, pregnancy >=22 weeks: Administer NACT.
[...]

[Answer]
Structured MDT recommendations covering: (1) Gynecologic Oncology:
NACT to prolong gestation; (2) Chemotherapy: Carboplatin + Paclitaxel
regimen with dose/schedule; (3) Obstetrics: Fetal monitoring plan,
planned cesarean at 34-36 weeks; (4) Neonatology: NICU readiness
and neonatal assessment protocol.
```

Figure 7: Representative examples for **L4: Decision Support** scenarios.

```
You are a professional evaluation expert. You need to evaluate the
quality of the "Predicted Answer" based on the following four core elements:

- **Dialogue History** (Contextual information)
- **Current Question** (The user's specific request)
- **Gold Standard Answer** (Verified high-quality reference answer)
- **Predicted Answer** (The answer to be evaluated)

### Scoring Criteria [Independent scoring, strict and rigorous]:

1. Accuracy [Positive Score, Max 100 points]:
   Evaluate whether the answer is correct, aligns with user intent,
   covers key information completely (no omissions or redundancies),
   and is logically sound. Scoring starts from the highest standard;
   any errors or deficiencies result in strict deductions.

2. Hallucination [Negative Penalty, Max deduction 25 points]:
   Evaluate whether the answer contains any factual errors, baseless
   speculations, or fabricated content. Zero tolerance policy:
   any confirmed hallucination results in severe deductions.

3. Readability [Negative Penalty, Max deduction 25 points]:
   Evaluate whether the language is fluent and natural, checks for
   improper language mixing (e.g., unnecessary English-Chinese mixing),
   and controllable formatting (severe deductions for format issues
   like piled-up line breaks that hinder semantic understanding).

(Total Score Formula = max(0, Accuracy - Hallucination - Readability))

> Note: The "Gold Standard Answer" has passed strict review and
  represents a high standard baseline.

---

### Dialogue History
{Insert Dialogue History}

### Current Question
{Insert Original Question}

### Gold Standard Answer (Reference)
assistant: {Insert Gold Answer}

### Predicted Answer (To be evaluated)
assistant: {Insert Predicted Answer}

---

Please output your evaluation results in the following structure:

### Evaluation Analysis
[Conduct item-by-item comparative analysis here]

### Predicted Answer Score
\boxed{Total Score}
```

Figure 8: The specific prompt template used for the calibrated expert scoring model (Type C Tasks).