

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

CSC5051/MDS5110/CSC6052 : Natural Language Processing

Lecture 9: Medical LLMs — Towards Generalist Medical AI

Fall 2026

Benyou Wang
School of Data Science

About Me

2018-2022, Phd, **University of Padova**

2022-~, The Chinese University of Hong Kong, Shenzhen (CUHKSZ)

Assistant professor in the School of Data Science and the School of Medicine, CUHK-Shenzhen, and the Shenzhen Loop Area Institute (SLAI).

Research:

- **Planning** of Agent (complex reasoning/math modeling/automatic theorem proof)
- **Perception** of Agent (audio,vision, time-series, etc.)
- **Human-agent interaction** (HCI, agent similation, brain-computer interface etc.)
- **Applications** of Agent (**Medical**/finnacial/legal/education)

Best papers:

- Best paper Honorable mention for SIGIR 2017
- NAACL 2019 Best Explainable NLP paper
- NLPCC 2022 Best Paper
- ICLR 2025 Finnacial AI 2025

Huggingface: https://huggingface.co/FreedomIntelligence

GitHub: https://github.com/FreedomIntelligence

Contents

Foundations — *why and what*

1 Why Medical AI?
Why a specific domain · why medicine · why now

2 The Story of HuatuoGPT
Data · two-stage training · evaluation · HuatuoGPT-II

3 What Is Generalist Medical AI?
Definition · language-centric · top-down vs. bottom-up · the ladder

Roadmap — *capabilities and challenges*

4 The Roadmap: Perception, Reasoning, Generation, Languages
Multimodal perception · long context · medical reasoning · generation · multilinguality

5 Challenges on the Way
Decentralized data · specialists vs. generalists · interpretability · benchmarking

Vision — *environments, applications, outlook*

6 Our Vision: From Models to Environments
Patients, world models, hospitals, cells, community evidence

7 Applications and Ecosystem
Regional deployment · education · skills · open source

8 Outlook
Five bets · AI entrepreneurship

The thesis of this lecture

Generalist Medical AI is the goal.

Trustworthy Medical AI is the way there.

Environments are the engine.

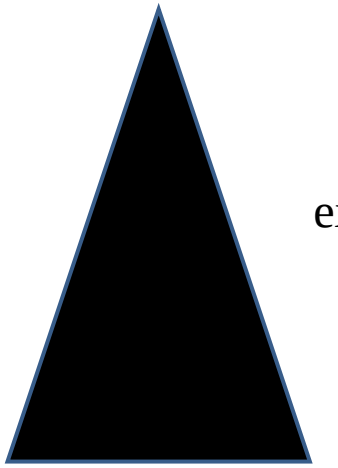
Standardized patients · patient trajectories · clinical world models · agentic hospitals · virtual cells · community evidence

1 Why Medical AI?

Medicine is where language models can create the most value — and where mistakes cost the most.

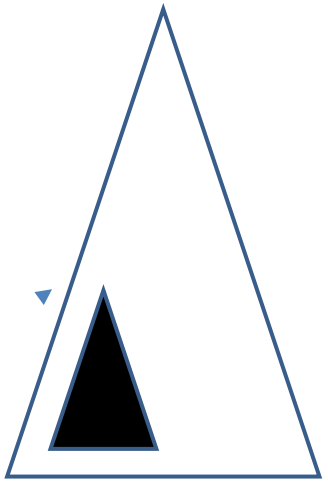
Why Work on a Specific Domain? (1) Competing with Giants

- In general domain, it is quite challenging to compete with giant companies. There are many companies working there, do not waste of resources on these over-competed domains.
- Medical AGI is Sandbox of AGI where the former needs less computing resources



expensive

AGI



Relatively cheaper

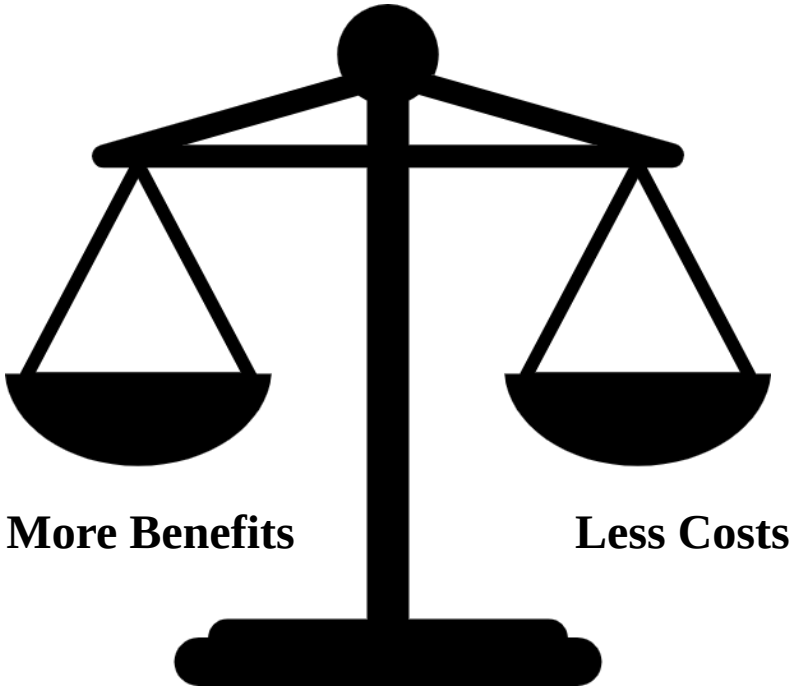
Medical AGI is the Sandbox of AGI

Why Work on a Specific Domain? (2) From General to Professional

LLMs are already good at general problems; the next step is professional problems.

Can an LLM do a task that takes an average professional five hours?

Why Work on a Specific Domain? (3) Benefits vs. Costs



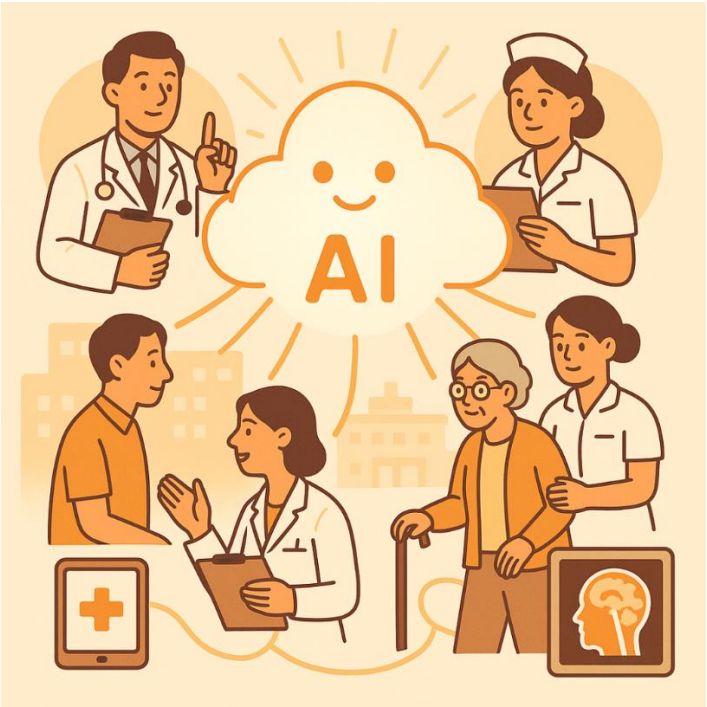
↑ **Benefits** > **Costs** ↓

Doing valuable applications **(LLM Applications) to** make our research more sustainable.

- Everyone needs **healthcare;** 5%-20% GPT for healthcare for most countries
- **One road one belt** (Arabic LLMs and multilingual LLMS)
- **AI reshapes Education** via Agent and better interactions

Why Medicine?

- Where
  - it benefits a **large population**
  - it is **challenging** enough
  - it could get sustainable **funding**
  - it allows public-welfare work as a university



Why Medical AI Now?

Demand

- Aging populations and chronic disease
- WHO projects a shortfall of ~10–11 million health workers by 2030
- Unequal access: experts concentrate in big cities

Capability

- AMIE (Nature 2025): in a blinded OSCE with patient actors, specialists rated it above primary-care physicians on 30 of 32 axes
- Reasoning models and multimodal models keep improving

Adoption

- OpenEvidence reports use by ~65% of US physicians (Apr 2026)
- Regional deployments in China, e.g. HuatuoGPT in Longgang (5M people)

The gap is no longer exam scores. It is trust, workflow integration and evidence of real-world benefit.

WHO Health Workforce (2016/2022 estimates) · Tu et al. Towards Conversational Diagnostic AI, Nature 2025 · NBC News on OpenEvidence (May 2026; company figures).

2 The Story of HuatuoGPT

Our first medical LLM taught us three lessons: data must look like doctors, knowledge must be injected, and evaluation is hard.

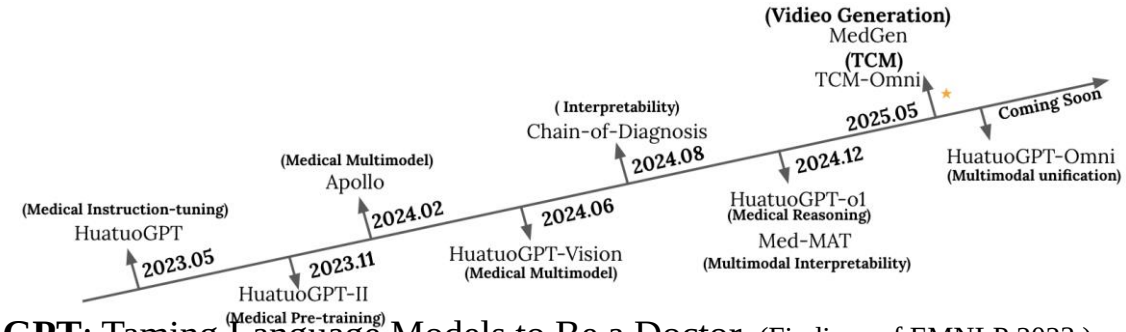
In this part

2.1 Building HuatuoGPT

2.2 How Do We Evaluate a Medical LLM?

2.3 HuatuoGPT-II: Injecting Medical Knowledge

Timeline of Our Medical LLMs



- 2023 May, **HuatuoGPT**: Taming Language Models to Be a Doctor. (Findings of EMNLP 2023.)
- 2023 May, Huatuo-26M: the largest medical QA **dataset.** (Findings of NAACL 2023).
- 2023 Nov, **HuatuoGPT-II**: One-stage Training for Medical Adaptation of LLMs. (COLM 2024).
- 2024 March, **Apollo: multilingual** medical LLMs for 6 languages
- 2024 June, **HuatuoGPT-Vision**: **Multimodal (visual)** Medical LLMs. (EMNLP 2024).
- 2024 Oct, HuatuoGPT **deployed** in 12 local hospitals
- 2024 Nov. **Apollo 2: multilingual** medical LLMs for 50 languages. (ICLR 2025)
- 2024 Dec, **HuatuoGPT-o1**: Towards Medical **Complex Reasoning** with LLMs. (Findings of ACL 2025).
- 2024 Dec, **HuatuoGPT-o1**: Towards Medical Complex Reasoning with LLMs
- 2025 June, MedGen: the medical video generation model
- 2025 August, ShizhenGPT : the first medical LLMs supporting full modalities for the traditional medicine

[1] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, **Benyou Wang**. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. https://arxiv.org/abs/2412.18925
[2] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, **Benyou Wang**. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. https://arxiv.org/abs/2406.19280. EMNLP 2024
[3] Junying Chen, Xidong Wang, Anningzhe Gao#, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, **Benyou Wang.** Huatuogpt-ii, one-stage training for medical adaption of llms. https://arxiv.org/abs/2311.09774. COLM 2024
[4] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, **Benyou Wang**, Haizhou Li. HuatuoGPT, towards Taming Language Model to Be a Doctor. https://arxiv.org/abs/2305.15075. Findings of EMNLP 2023
[5] Zhihong Chen, Shizhe Diao, **Benyou Wang**, Guanbin Li, and Xiang Wan. "Towards Unifying Medical Vision-and-Language Pre-training via Soft Prompts". **ICCV 2023**
[6] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, **Benyou Wang**, Haizhou Li .CMB: A Comprehensive Medical Benchmark in Chinese**. NAACL 2024**

Starting from Data

Composition	# Pairs	Len(Q)	Len(A)
Huatuo-26M Train	26,239,047	44.6	120.7
Huatuo-26M Test	265,041	44.6	120.6
Data source:			
Consultant records	25,341,578	46.0	117.3
Encyclopedias	364,066	11.5	540.4
Knowledge bases	798,444	15.8	35.9
All	26,504,088	44.6	120.7

Table 3: Basic statistics of Huatuo-26M.

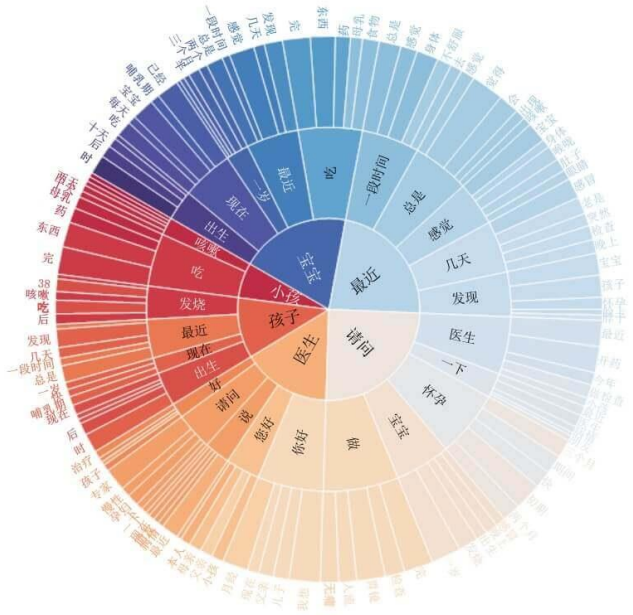


Figure 1: Distribution of patient-focused questions. We do the analysis heuristically, counting from the first meaningful phrase of the question. We present the relative distribution of these recurring problems and their subsequent distributions.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, Benyou Wang. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. Findings of NAACL 2025.

Huatuo-26M: The Largest Chinese Medical QA Dataset

Dataset	Lang	Domain	Source	#Q
MedHop (Welbl et al., 2018)	English	Medical	MEDLINE	2.5K
BiQA (Lamurias et al., 2020)	English	Medical	Online Medical forum	7.4K
HealthQA (Zhu et al., 2019)	English	Medical	Medical-services website	7.5K
MASH-QA (Zhu et al., 2020)	English	Medical	Medical article website	35K
MedQuAD (Ben Abacha and Demner-Fushman, 2019)	English	Medical	U.S. National Institutes of Health (NIH)	47K
ChiMed (Tian et al., 2019)	Chinese	Medical	Online Medical forum	47K
MedQA (Jin et al., 2020)	EN&CH	Medical	Medical Exam	60K
webMedQA (He et al., 2019)	Chinese	Medical	Medical consultancy websites	63K
CliCR (Šuster and Daelemans, 2018)	English	Medical	Clinical case reports	100K
cMedQA2 (Zhang et al., 2018)	Chinese	Medical	Online Medical forum	108K
Huatuo-26M	**Chinese**	**Medical**	**Consultation records, Encyclopedia, KBs**	**26M**
TriviaQA (Joshi et al., 2017)	English	General	Trivia	96K
HotpotQA (Yang et al., 2018)	English	General	Wikipedia	113K
SQuAD (Rajpurkar et al., 2016)	English	General	Wikipedia	158K
DuReader (He et al., 2017)	Chinese	General	Web search	200K
Natural Questions (Kwiatkowski et al., 2019)	English	General	Wikipedia	323K
MS MARCO (Nguyen et al., 2016)	English	General	Web search	1.0M
CNN/Daily Mail (See et al., 2017)	English	General	News	1.3M
PAQ (Lewis et al., 2021)	English	General	Wikipedia	65M

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, Benyou Wang. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. https://arxiv.org/abs/2305.01526

HuatuoGPT: Training Recipe

Pre-training backbone

- **Bloom-7B**
- **Ziya-13B**
- **Baichuan-7B/13B**

- Pre-training data (5B tokens)
  - Textbooks
  - Data from internet
  - Medical guidelines

Finetuning

- Mixture of instruction-following data and conversation
- Careful data engineering

RLHF

- ChatGPT to provide feedbacks
- Doctors’ answers as reference

Motivation: Doctors vs. ChatGPT

- With rich format (e.g., bullets)
- Usually long
- Well-organized structure
- Few follow-up questions
- **Reading**-friendly

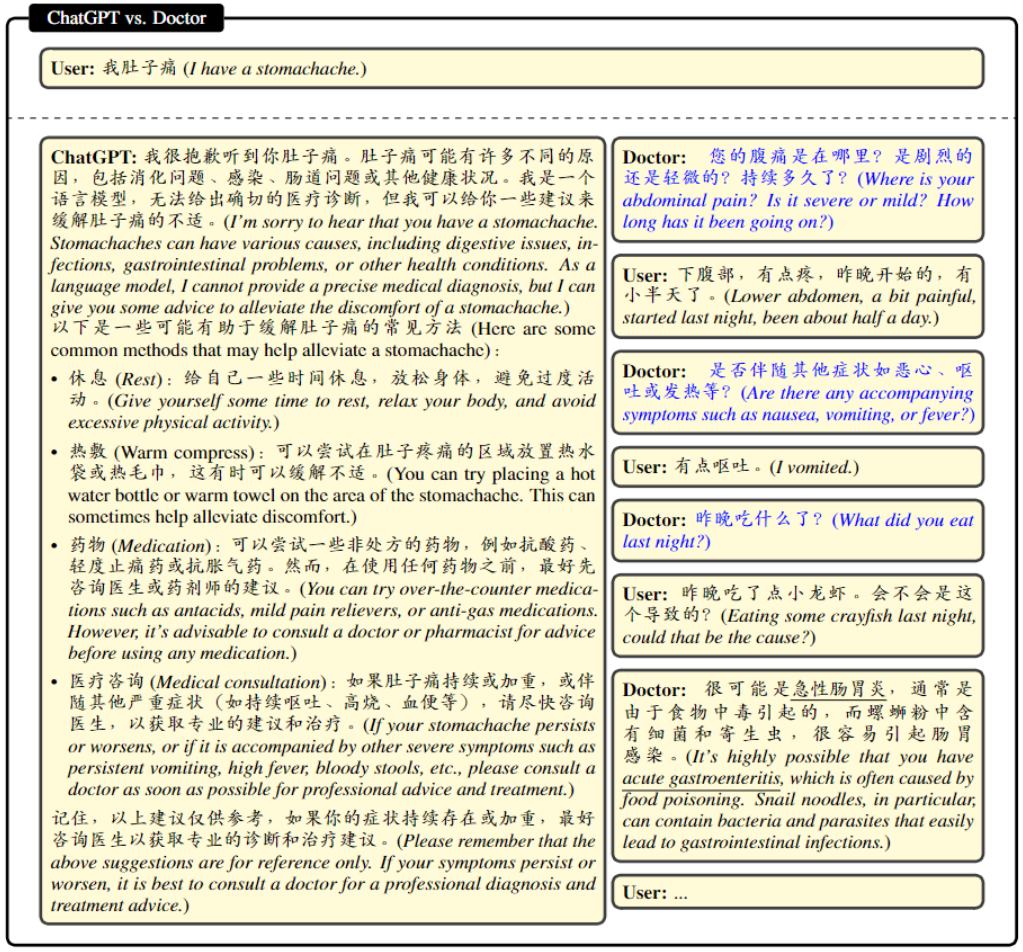
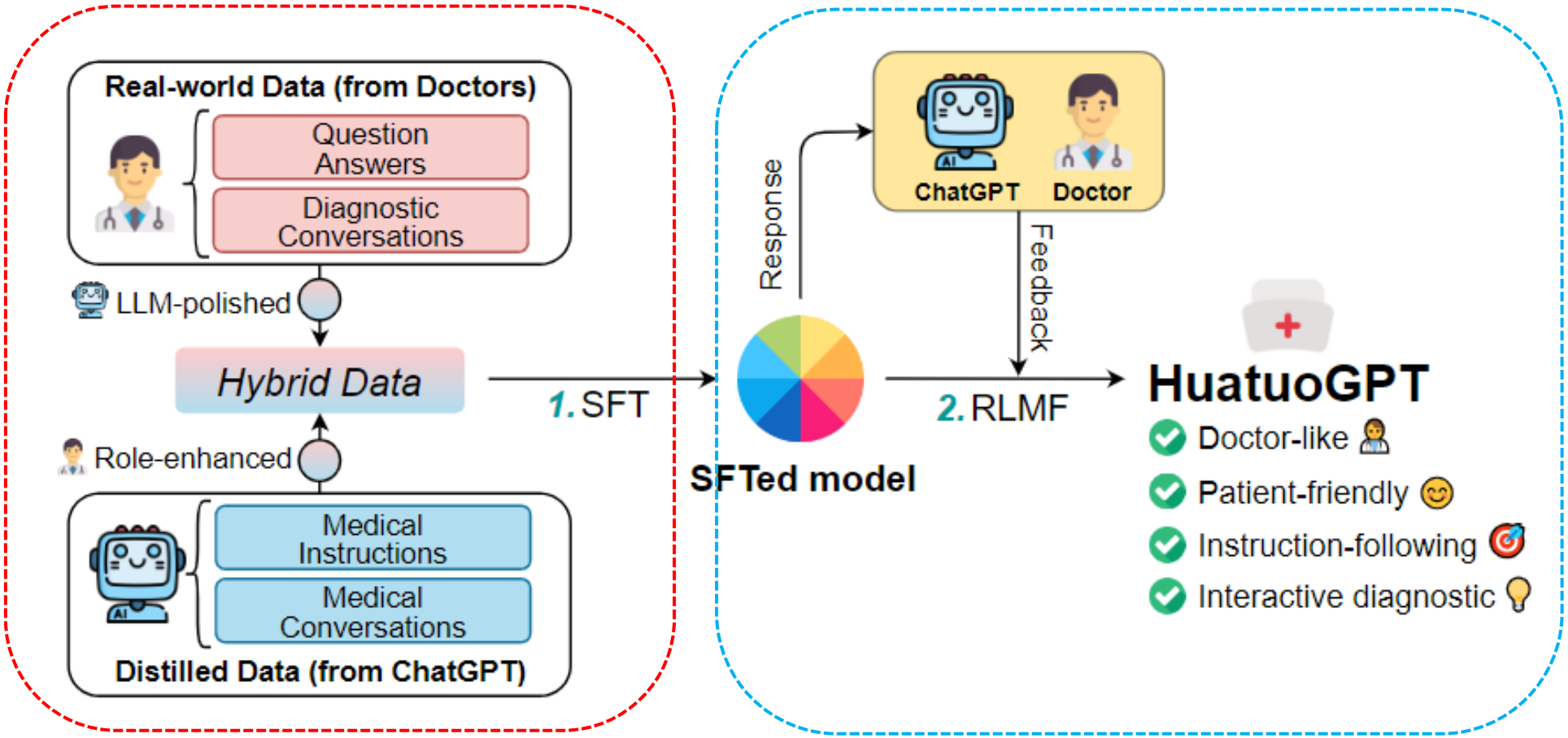


Figure 1: Example of ChatGPT responses (left) and doctor responses (right) in Medical Consultation Dialogue, where texts are *translated* from Chinese to English. Questions raised by doctors are in blue, and medical diagnoses are underlined. Note that **ChatGPT usually does not raise questions in response to patients or provide medical diagnoses like doctors.**

- Usually short
- Without rich format
- Diagnosis in multi turns
  - (e.g. follow-up questions)
- **Orally**-friendly

- Doctors' replies often jump to conclusions without intermediate steps, so the reasoning is lost
  (e.g., "You are pregnant, take this drug" — with no explanation)
- Training only on doctors' data hurts instruction following.
  e.g., the user asks three questions and the doctor answers only one of them.

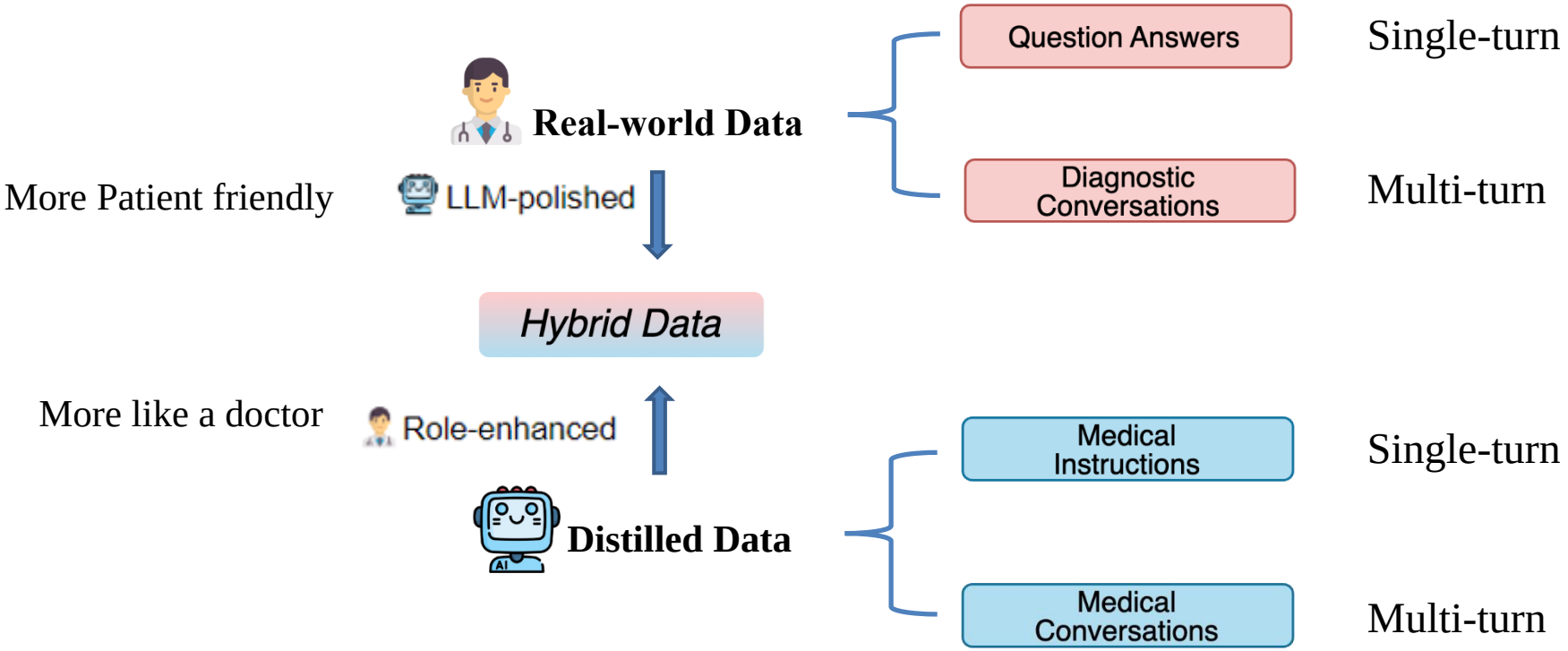
HuatuoGPT: Two-Stage Training



SFT with Hybrid Data

RL from Mixed Feedback

Stage 1: SFT with Hybrid Data



Hybrid Data (1): Polish Doctors' Answers with an LLM



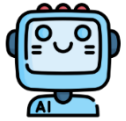
Question Answers

Single-turn QA

<Patient question>: ${Patient_Question}
<Reference reply>: ${Doctor_Response}
You are HuatuoGPT. Using the reference reply, answer the patient's question like a doctor, in a gentle and caring tone, with more detail and more help. If a diagnosis is not yet possible, ask the patient for more information.
<HuatuoGPT reply>:

ChatGPT answers questions based on doctor‘s responses:
making the answers are uniform, fluent and instruction-following.

Hybrid Data (2): Let the LLM Play a Doctor



Medical Instructions

Medical instructions generated with **self-instruction.**

System Prompt

Design 20 different <role, instruction, input> triples: the role, an instruction that role would give GPT to be more productive, and the input for that instruction. Requirements:
1. Roles must relate to healthcare and can be specific, e.g., “respiratory physician”.
2. Instructions must be diverse (types and verbs), doable by a text-only LLM, 1–2 sentences long, and may use a placeholder filled by the input.
3. Inputs must be concrete, real content (not a link or file name), ideally under 200 characters.
4. No duplicates; the instruction is required, the input may be empty. The 20 triples are:

System Prompt

Suppose you are an experienced [Medical Role] who answers patients patiently and thoroughly, in a gentle tone. If the patient gives too little information, you ask follow-up questions; at the end you add extra advice. If the patient asks:
{Question}
then you would answer:

1. Generate medical instructions **with roles via self-instruction**.

2. **mimics a medical role** to answers questions of the medical instructions.

Hybrid SFT Data in Numbers

Sources	Type	Scale
From the ChatGPT	Instruction	61400
	Conversation	68888
From doctors	Instruction	69768
	Conversation	25986
Total		226042

Training Loss

1. Mix single-round conversations and multi-round conversations in the same SFT process.
2. Do not split multi-round conversations in to several parts.

Single-turn:

System_Prompt + <Patient>: [User_Input] + <HuatuoGPT>: [Huatuo_Response]</s>

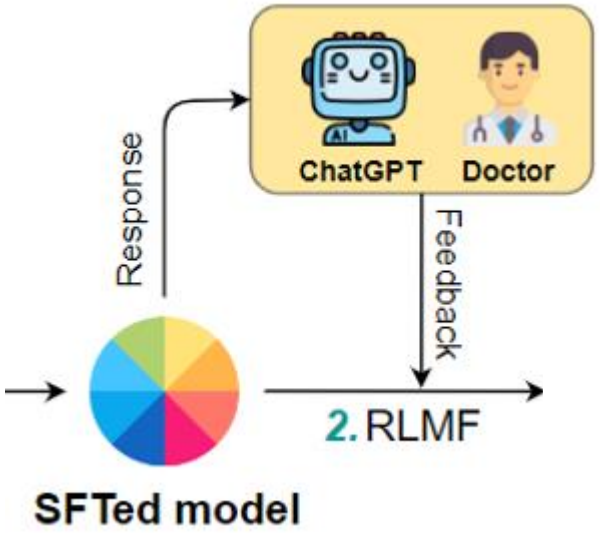
Loss

Multi-turn:

System_Prompt + <Patient>: [User_Input] + <HuatuoGPT>: [Huatuo_Response]</s> <Patient>: [User_Input] + <HuatuoGPT>: [Huatuo_Response]</s> …

Loss

Stage 2: RL with Mixed Feedback (RLMF)



RLMF

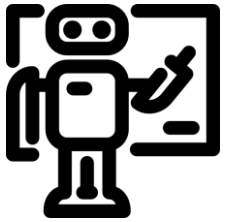
RLHF



High-Costing

Time-consuming

RLAIF

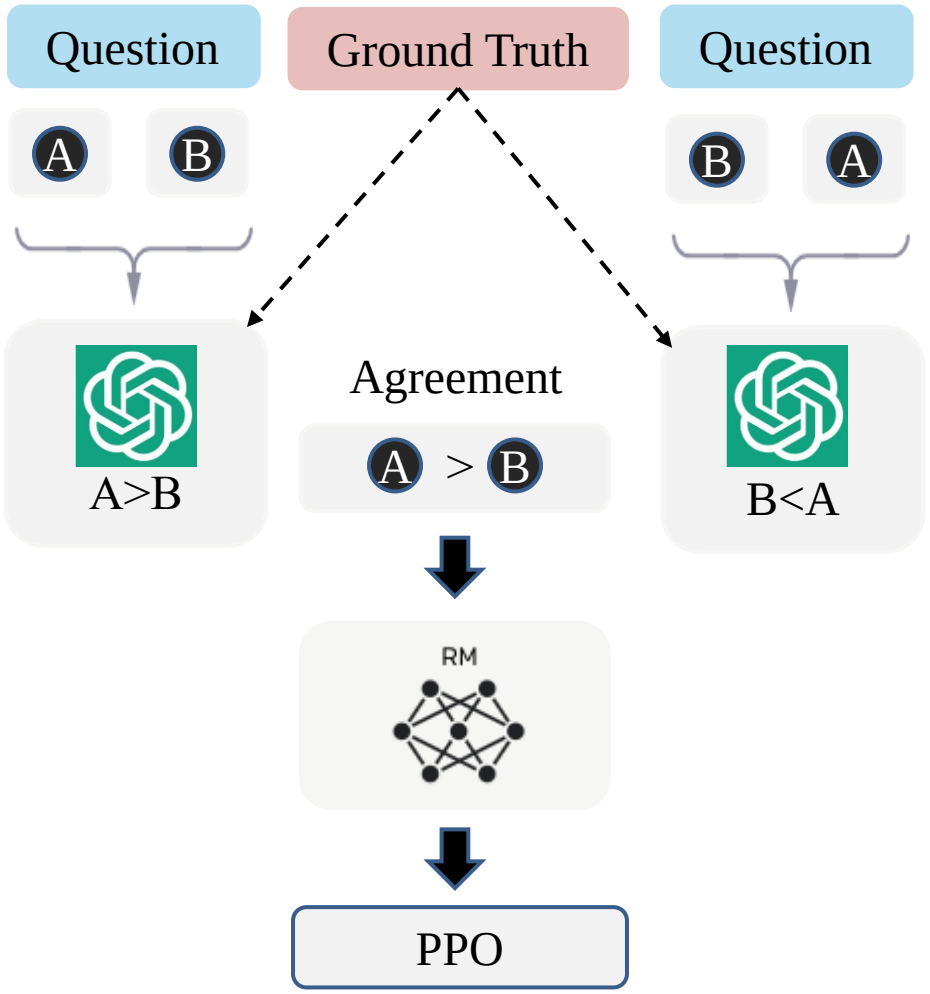


Medical inaccuracy

Unstable

Ouyang et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
Bai et al. Constitutional AI: Harmlessness from AI Feedback[J]. arXiv preprint arXiv:2212.08073, 2022.

RLMF: Reward Model Training



Here is a conversation history:
[History]
${History}
[End of History]
Here is the final question and the standard answer:
[Question]
${Query}
[End of question]
[Standard answer]
${Doctor_response}
[End of standard answer]
Based on the conversation history, user question, and standard answer, please rate the following two AI responses on a scale of 1 to 10,**considering accuracy, conciseness, and similarity to the standard answer.**
Please provide the ratings in the following format: "Rating A: [score];Rating B: [score]".
[Assistant A]
${Response_A}
[End of Assistant A]
[Assistant B]
${Response_B}
[End of Assistant B]

Ouyang et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
Bai et al. Constitutional AI: Harmlessness from AI Feedback[J]. arXiv preprint arXiv:2212.08073, 2022.

RLMF: Reward Curve



runs.summary["samples"]

	prompt	output	reward
195	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回	您好，根据您的描述，孩子可能患有食道裂孔疝并且出现呕吐症状。这种情况可能是由于消化不良引起的胃肠功能紊乱所致。	10.482
131	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回	好，根据您的描述，您媳妇患有慢性肾脏疾病已经四年了，而且B超检查显示双肾已经有萎缩的现象。这是一种比较常见的病情表现，需要引起重视	10.109
22	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回		
275	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回		
458	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回		
454	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回		
43	一位用户和智能医疗大模型HuatuoGPT之间的对话。对于用户的医疗问诊，HuatuoGPT给出准确的、详细的、温暖的指导建议。对于用户的指令问题，HuatuoGPT给出有益的、详细的、有礼貌的回		

String preview

您好，根据您的描述，您媳妇患有慢性肾脏疾病已经四年了，而且B超检查显示双肾已经有萎缩的现象。这是一种比较常见的病情表现，需要引起重视。

治疗方案应该从改善血液动力学角度考虑。目前常用的药物包括扩血管药如硝酸甘油等以及一些活血化瘀的中成药，如丹参片或复方丹参注射液等等。这些药物可以增加尿量，促进心排出量的恢复，从而达到改善循环的目的。此外，也可以使用利尿剂如螺内酯来减少水钠潴留及血压下降的作用。需要注意的是，对高血压性心脏病的防治措施应注意控制高血压和高血脂症。同时，积极纠正贫血症状也是必要的。

除此之外，还需要定期进行肾功能检查，以便及时发现并处理任何异常情况。如果出现严重的肾脏损害，可能导致不可逆转的心脏损伤，最终发生心力衰竭甚至猝死。因此，一旦确诊为CKD2期以上，就需要接受更加严格的管理和治疗计划，以避免不良后果的发生。

希望我的回复能够帮助到您，如果您还有其他问题或者需要更详细的解答，请随时向我提出。祝您和您的家人健康！

Training Details

Our model is implemented in **PyTorch** using the **Accelerate** and **trlx** packages with LLaMA as the base architecture.
initialize the model parameters using **Ziya-LLaMA-13B-Pretrain**.

We leverage **ZeRO** strategy to distribute the model across 8 * A100 (80G) GPUs for training.

SFT	
learning rate	5e-5
batch size	128
maximum context length	2048

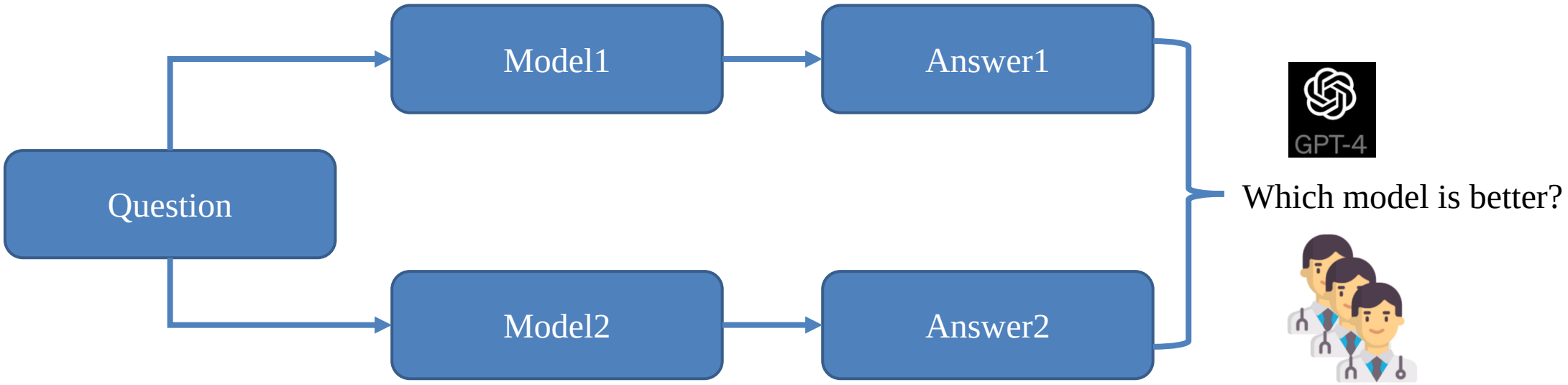
RLMF	
process encompasses	16000
learning rate	1e-6
rollout size	64
chunk size	8
epoch	4
KL divergence coefficient	0.1

https://huggingface.co/docs/accelerate/index
https://github.com/CarperAI/trlx

2.2 How Do We Evaluate a Medical LLM?

Leaderboards saturate and leak; doctors disagree; real medicine is interactive. Evaluation became a research problem in itself.

Evaluation: Single-Turn Questions



Chiang et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% ChatGPT quality.

Single-Turn Questions in 10 Intents

Single-turn Questions in 10 intents from KUAKE-QIC of CBLUE

Diagnosis: given symptoms, find possible causes
- Why do I feel weak every morning lately?

Etiology: explain why a known disease occurs
- How does nasopharyngeal carcinoma develop?

Treatment: plans for a disease or symptom (tests, surgery, drugs, behavior)
- How can I remove a burn scar?

Care advice: which department or test to go to
- Which tests should a diabetic patient take?

Lab interpretation: what a test value means
- What does an elevated hs-CRP mean?

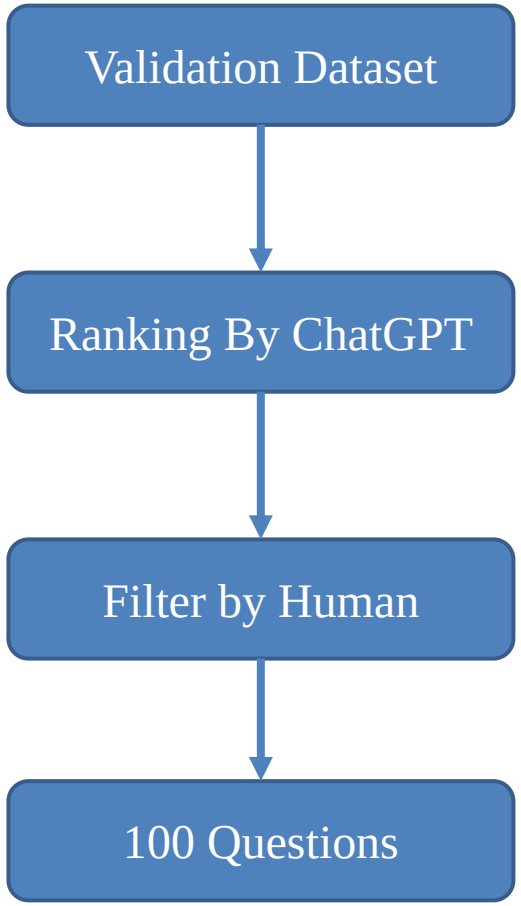
Disease description: curability, symptoms, presentation
- What are the early symptoms of external hemorrhoids?

Consequences: harms of diseases, drugs or food; outcomes with or without treatment
- What happens with potassium deficiency?

Precautions: what patients should watch for, including diet
- What should asthma patients be careful about?

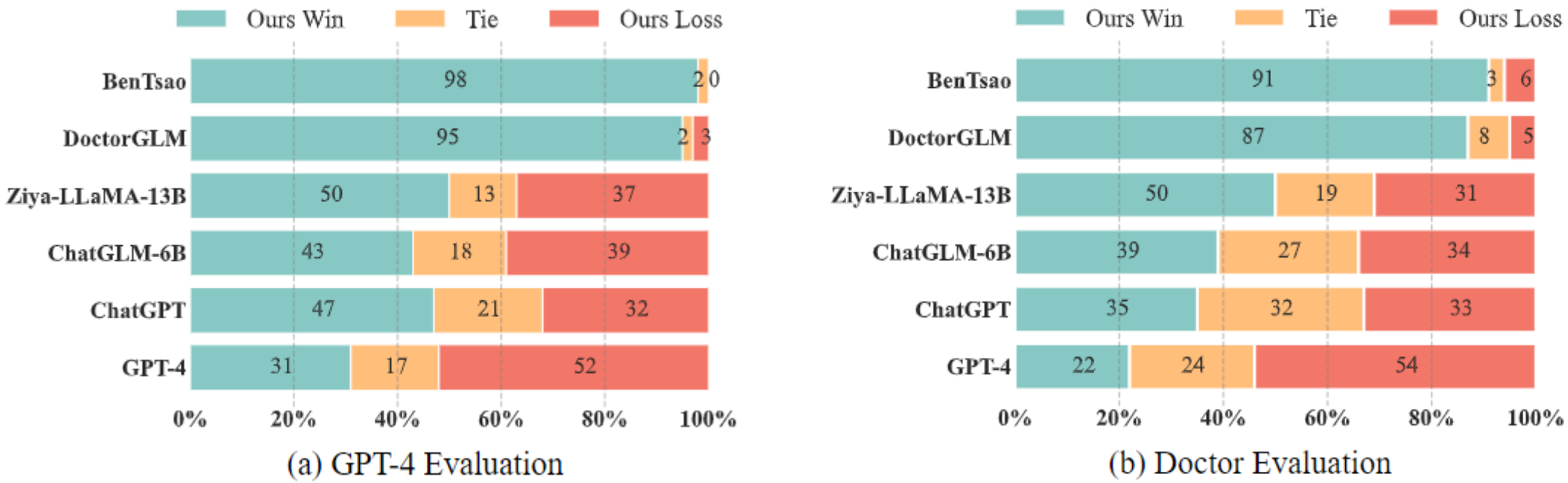
Effects: benefits and side effects of foods or drugs
- What are the effects of Wuji Baifeng pills?

Costs: of diseases, surgery, drugs or tests
- How much does metformin cost?



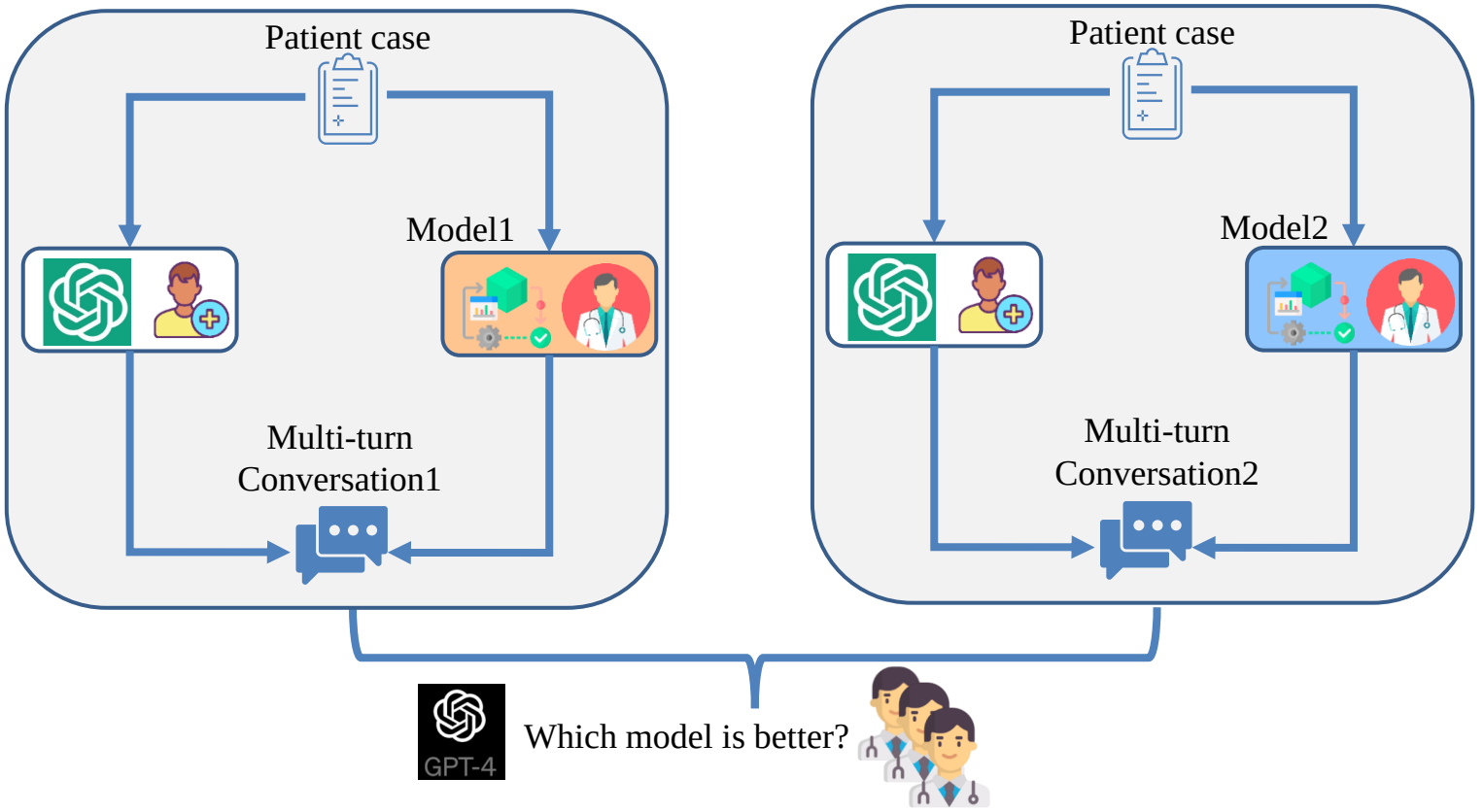
Zhang et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark[J]. arXiv preprint arXiv:2106.08087, 2021.

Single-Turn Results

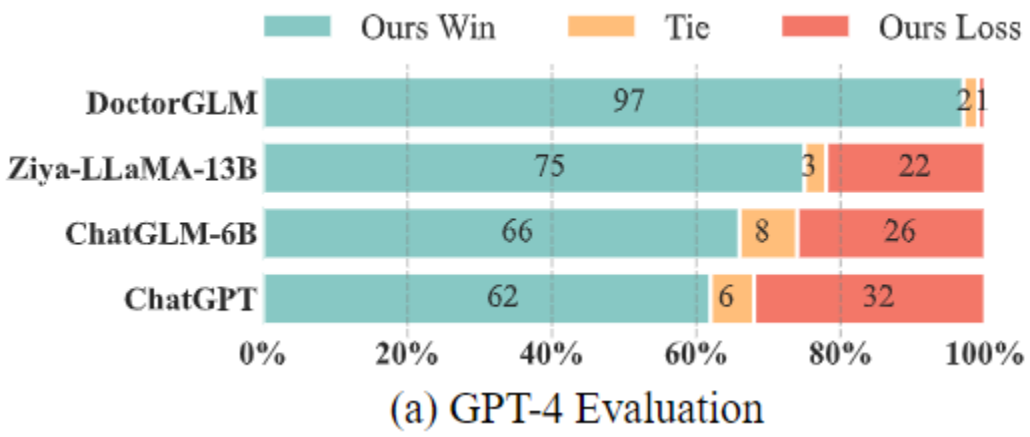


Response Comparison of HuatuoGPT with Other Baselines on the Single-turn Question

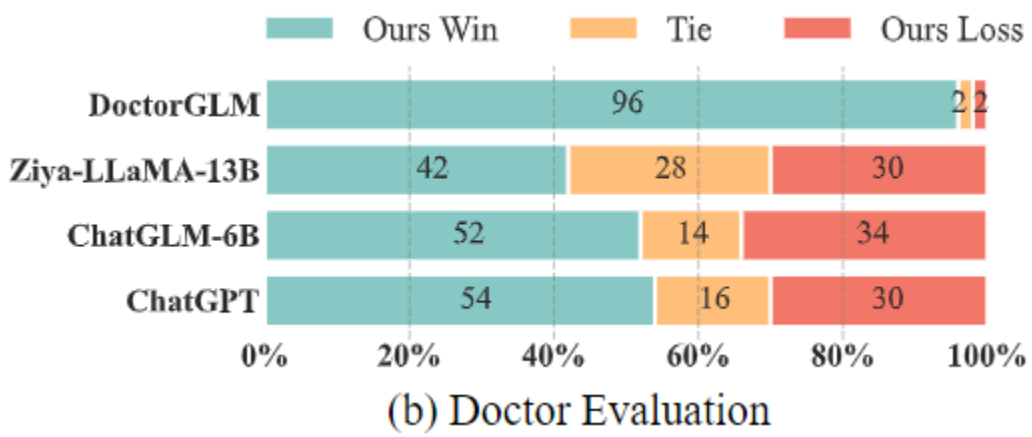
Evaluation: Multi-Turn Conversations



Multi-Turn Results



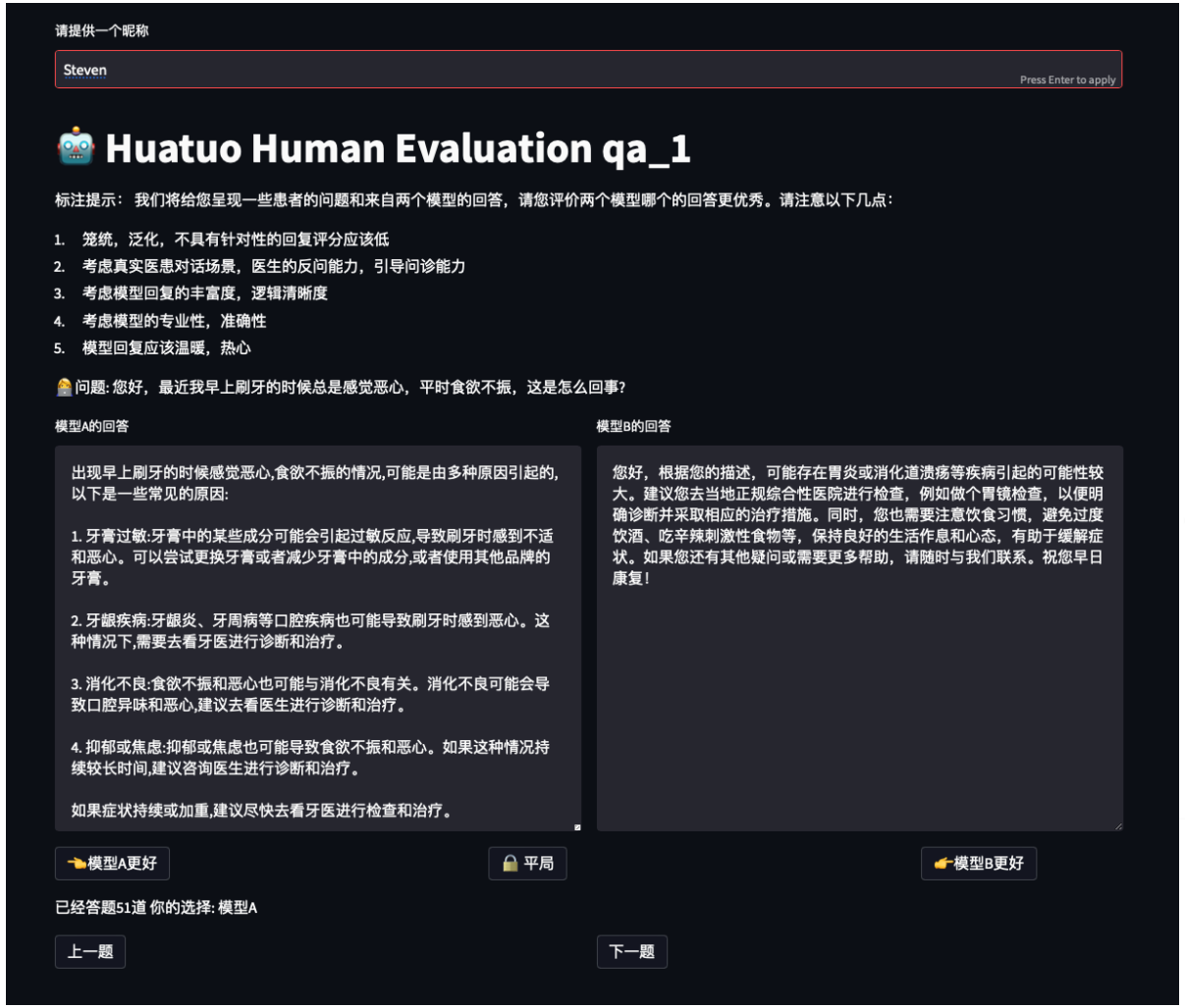
(a) GPT-4 Evaluation

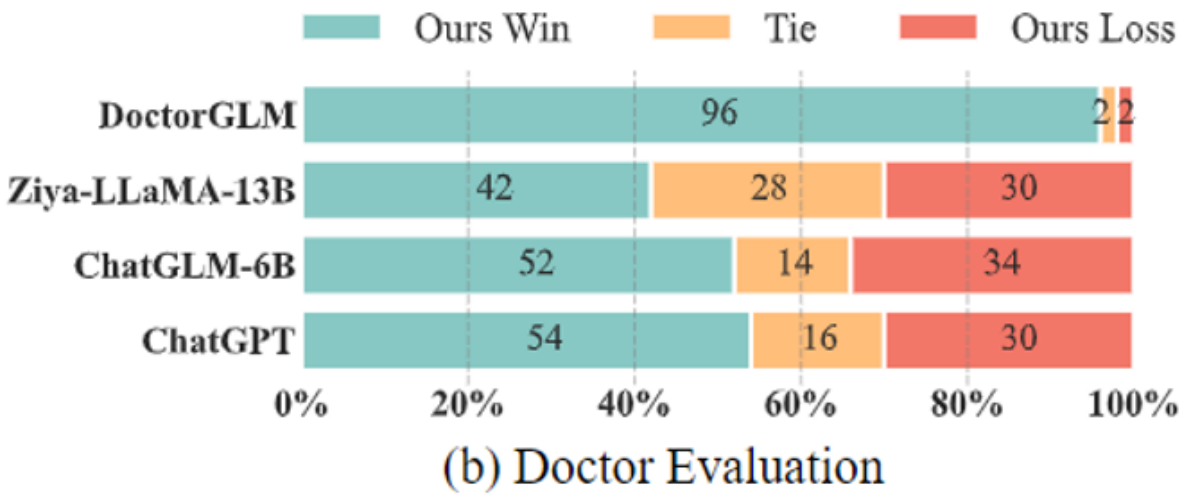


(b) Doctor Evaluation

Response Comparison of HuatuoGPT with Other Baselines on the Multi-turn Conversations

Human Evaluation





(b) Doctor Evaluation

Medical QA Benchmarks

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
cMedQA2	T5 (fine-tuned)	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.41	0.52
	DoctorGLM	13.51	7.10	3.72	2.00	5.11	22.78	5.68	12.22	**0.85**	0.96
	ChatGPT	19.21	7.43	3.14	1.24	5.06	20.13	3.10	12.57	0.69	**0.99**
	ChatGLM-6B	24.90	12.74	6.99	3.87	8.49	28.52	7.19	**18.21**	0.68	**0.99**
	Ziya-LLaMA-13B	27.03	13.87	7.48	4.09	7.77	28.24	7.10	14.81	0.78	0.93
	HuatuoGPT	**27.39**	**14.38**	**8.06**	**4.55**	**8.52**	**29.26**	**8.02**	15.46	0.74	0.93
webMedQA	T5 (fine-tuned)	21.42	13.79	10.06	7.38	8.94	31.00	13.85	25.78	0.37	0.46
	DoctorGLM	9.91	5.20	2.78	1.54	4.67	23.01	5.68	11.96	**0.84**	**0.95**
	ChatGPT	18.06	6.74	2.73	1.09	4.71	20.01	2.81	12.58	0.65	0.87
	ChatGLM-6B	23.42	12.10	6.73	3.83	**8.04**	**28.30**	6.87	**18.49**	0.63	0.87
	-LLaMa-13B	22.16	11.70	6.53	3.74	6.91	27.41	6.80	13.52	0.76	0.93
	HuatuoGPT	**24.85**	**13.42**	**7.72**	**4.51**	7.50	**28.30**	**7.72**	14.50	0.73	0.93
Huatuo-26M	T5 (fine-tuned)	26.63	16.74	11.77	8.46	11.38	33.21	13.26	24.85	0.51	0.68
	DoctorGLM	11.50	6.00	3.14	1.69	4.65	22.39	5.47	12.14	**0.85**	**0.96**
	ChatGPT	18.44	6.95	2.87	1.13	4.87	19.60	2.82	12.46	0.69	0.89
	ChatGLM-6B	24.46	12.75	7.20	4.13	**8.50**	28.44	7.31	**18.58**	0.67	0.89
	Ziya-LLaMA-13B	25.58	13.39	7.46	4.24	7.30	28.14	7.18	14.78	0.77	0.93
	HuatuoGPT	**27.42**	**14.84**	**8.54**	**4.96**	8.01	**29.16**	**8.29**	15.84	0.74	0.93

Benchmark on Chinese medical QA dataset

Zhang et al. 2018. Multi-scale attentive interaction networks for Chinese medical question answer selection. IEEE Access, 6:74061–74071.

He et al. 2019. Applying deep matching networks to Chinese medical question answering: a study and a dataset. BMC medical informatics and decision making, 19(2):91–100.
Li et al. 2023. Huatuo-26M, a large-scale Chinese medical QA dataset. arXiv preprint arXiv:2305.01526.

CMB: A Comprehensive Medical Benchmark in Chinese

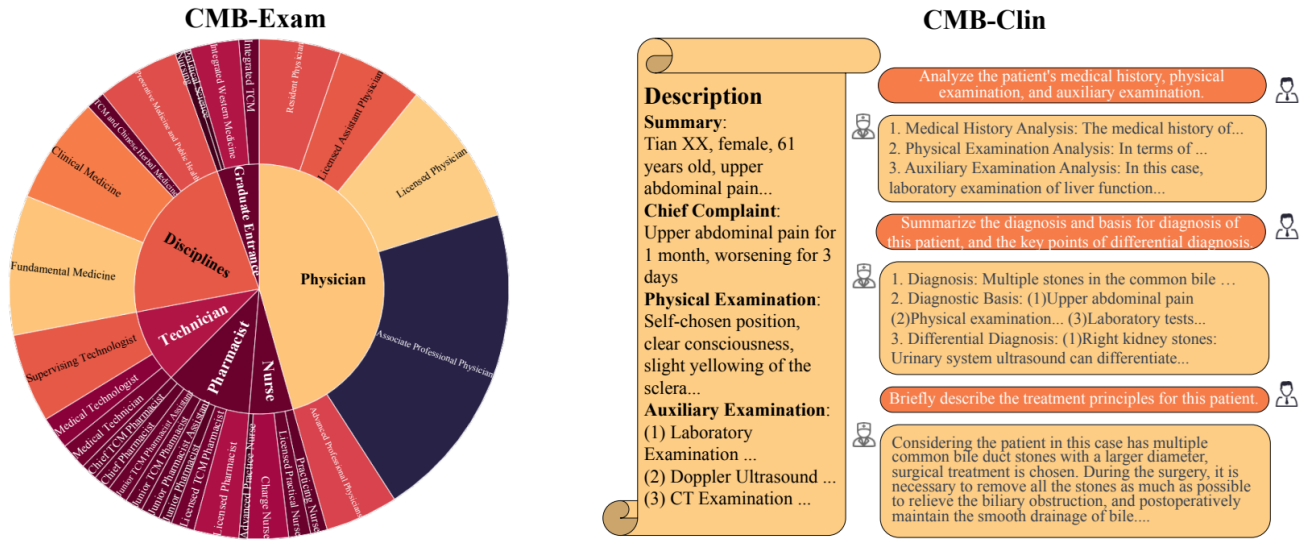
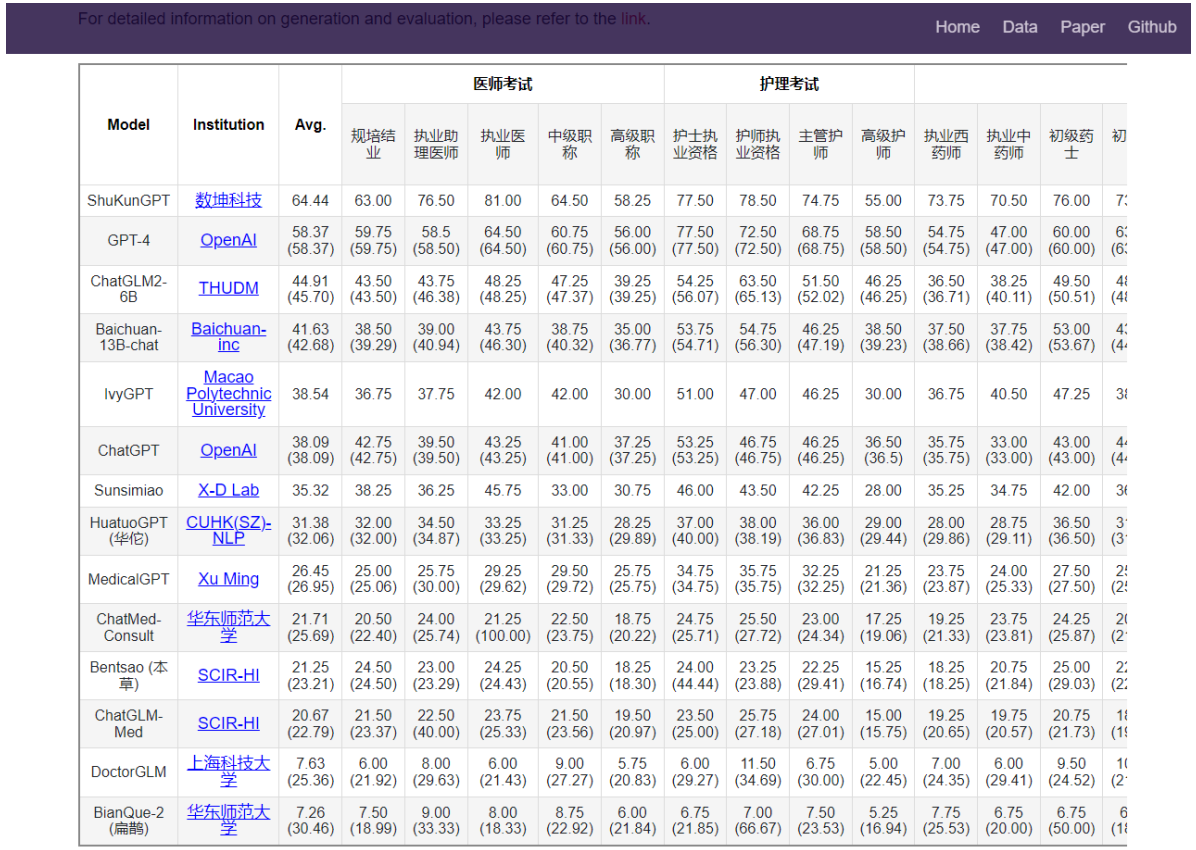


Figure 1: Components of the CMB dataset. Left: The structure of CMB-Exam, consisting of multiple-choice and multiple-answer questions. Right: an example of CMB-Clin. Each example consists of a description and a multi-turn conversation.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang*, Haizhou Li .**CMB: A Comprehensive Medical Benchmark in Chinese. NAACL 2024**

CMB Online Leaderboard

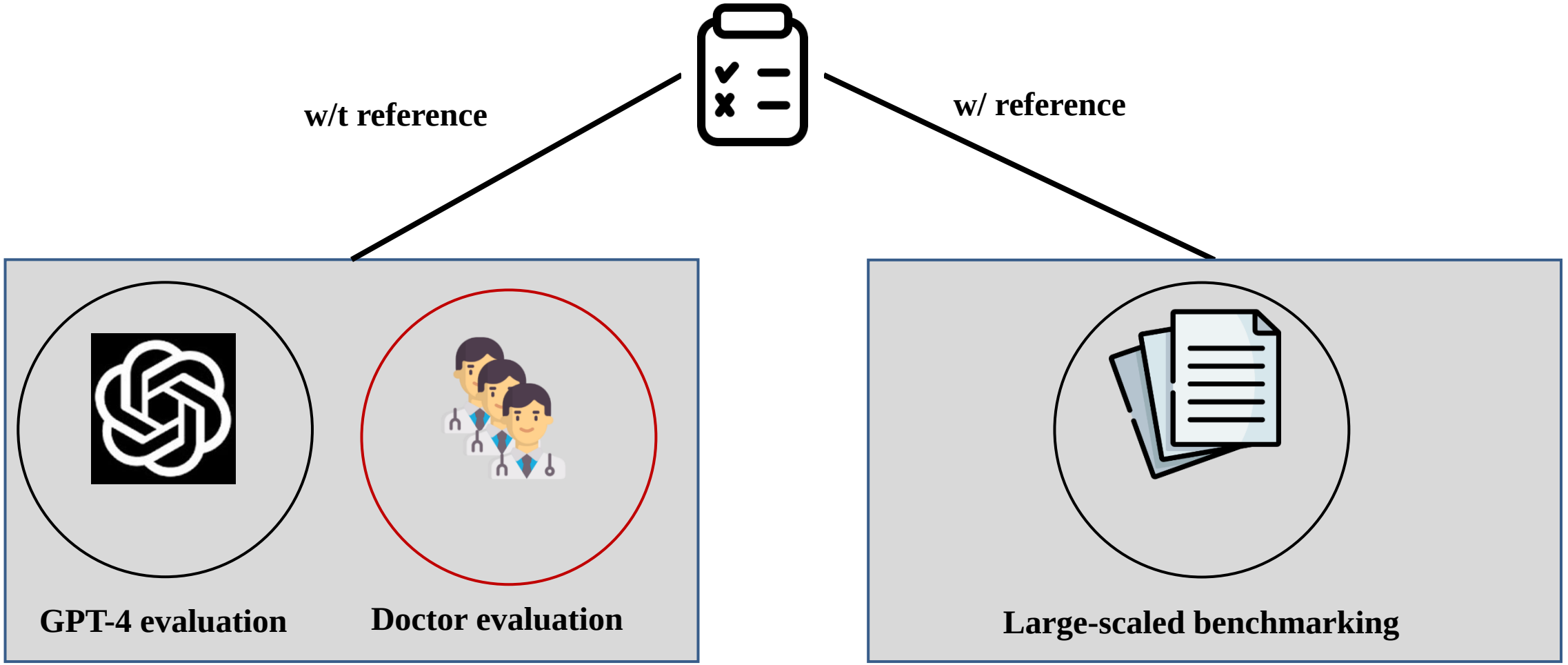


Model	Institution	Avg.	医师考试					护理考试								
			规培结业	执业助理医师	执业医师	中级职称	高级职称	护士执业资格	护师执业资格	主管护师	高级护师	执业西药师	执业中药师	初级药士	初
ShuKunGPT	数坤科技	64.44	63.00	76.50	81.00	64.50	58.25	77.50	78.50	74.75	55.00	73.75	70.50	76.00	7
GPT-4	OpenAI	58.37 (58.37)	59.75 (59.75)	58.5 (58.50)	64.50 (64.50)	60.75 (60.75)	56.00 (56.00)	77.50 (77.50)	72.50 (72.50)	68.75 (68.75)	58.50 (58.50)	54.75 (54.75)	47.00 (47.00)	60.00 (60.00)	6 (6
ChatGLM2-6B	THUDM	44.91 (45.70)	43.50 (43.50)	43.75 (46.38)	48.25 (48.25)	47.25 (47.37)	39.25 (39.25)	54.25 (56.07)	63.50 (65.13)	51.50 (52.02)	46.25 (46.25)	36.50 (36.71)	38.25 (40.11)	49.50 (50.51)	4 (4
Baichuan-13B-chat	Baichuan-inc	41.63 (42.68)	38.50 (39.29)	39.00 (40.94)	43.75 (46.30)	38.75 (40.32)	35.00 (36.77)	53.75 (54.71)	54.75 (56.30)	46.25 (47.19)	38.50 (39.23)	37.50 (38.66)	37.75 (38.42)	53.00 (53.67)	4 (4
IvyGPT	Macao Polytechnic University	38.54	36.75	37.75	42.00	42.00	30.00	51.00	47.00	46.25	30.00	36.75	40.50	47.25	3
ChatGPT	OpenAI	38.09 (38.09)	42.75 (42.75)	39.50 (39.50)	43.25 (43.25)	41.00 (41.00)	37.25 (37.25)	53.25 (53.25)	46.75 (46.75)	46.25 (46.25)	36.50 (36.5)	35.75 (35.75)	33.00 (33.00)	43.00 (43.00)	4 (4
Sunsimiao	X-D Lab	35.32	38.25	36.25	45.75	33.00	30.75	46.00	43.50	42.25	28.00	35.25	34.75	42.00	3
HuatuoGPT (华佗)	CUHK(SZ)-NLP	31.38 (32.06)	32.00 (32.00)	34.50 (34.87)	33.25 (33.25)	31.25 (31.33)	28.25 (29.89)	37.00 (40.00)	38.00 (38.19)	36.00 (36.83)	29.00 (29.44)	28.00 (29.86)	28.75 (29.11)	36.50 (36.50)	3 (3
MedicalGPT	Xu Ming	26.45 (26.95)	25.00 (25.06)	25.75 (30.00)	29.25 (29.62)	29.50 (29.72)	25.75 (25.75)	34.75 (34.75)	35.75 (35.75)	32.25 (32.25)	21.25 (21.36)	23.75 (23.87)	24.00 (25.33)	27.50 (27.50)	2 (2
ChatMed-Consult	华东师范大学	21.71 (25.69)	20.50 (22.40)	24.00 (25.74)	21.25 (100.00)	22.50 (23.75)	18.75 (20.22)	24.75 (25.71)	25.50 (27.72)	23.00 (24.34)	17.25 (19.06)	19.25 (21.33)	23.75 (23.81)	24.25 (25.87)	2 (2
Bentsao (本草)	SCIR-HI	21.25 (23.21)	24.50 (24.50)	23.00 (23.29)	24.25 (24.43)	20.50 (20.55)	18.25 (18.30)	24.00 (44.44)	23.25 (23.88)	22.25 (29.41)	15.25 (16.74)	18.25 (18.25)	20.75 (21.84)	25.00 (29.03)	2 (2
ChatGLM-Med	SCIR-HI	20.67 (22.79)	21.50 (23.37)	22.50 (40.00)	23.75 (25.33)	21.50 (23.56)	19.50 (20.97)	23.50 (25.00)	25.75 (27.18)	24.00 (27.01)	15.00 (15.75)	19.25 (20.65)	19.75 (20.57)	20.75 (21.73)	1 (1
DoctorGLM	上海科技大学	7.63 (25.36)	6.00 (21.92)	8.00 (29.63)	6.00 (21.43)	9.00 (27.27)	5.75 (20.83)	6.00 (29.27)	11.50 (34.69)	6.75 (30.00)	5.00 (22.45)	7.00 (24.35)	6.00 (29.41)	9.50 (24.52)	1 (2
BianQue-2 (扁鹊)	华东师范大学	7.26 (30.46)	7.50 (18.99)	9.00 (33.33)	8.00 (18.33)	8.75 (22.92)	6.00 (21.84)	6.75 (21.85)	7.00 (66.67)	7.50 (23.53)	5.25 (16.94)	7.75 (25.53)	6.75 (20.00)	6.75 (50.00)	6 (1

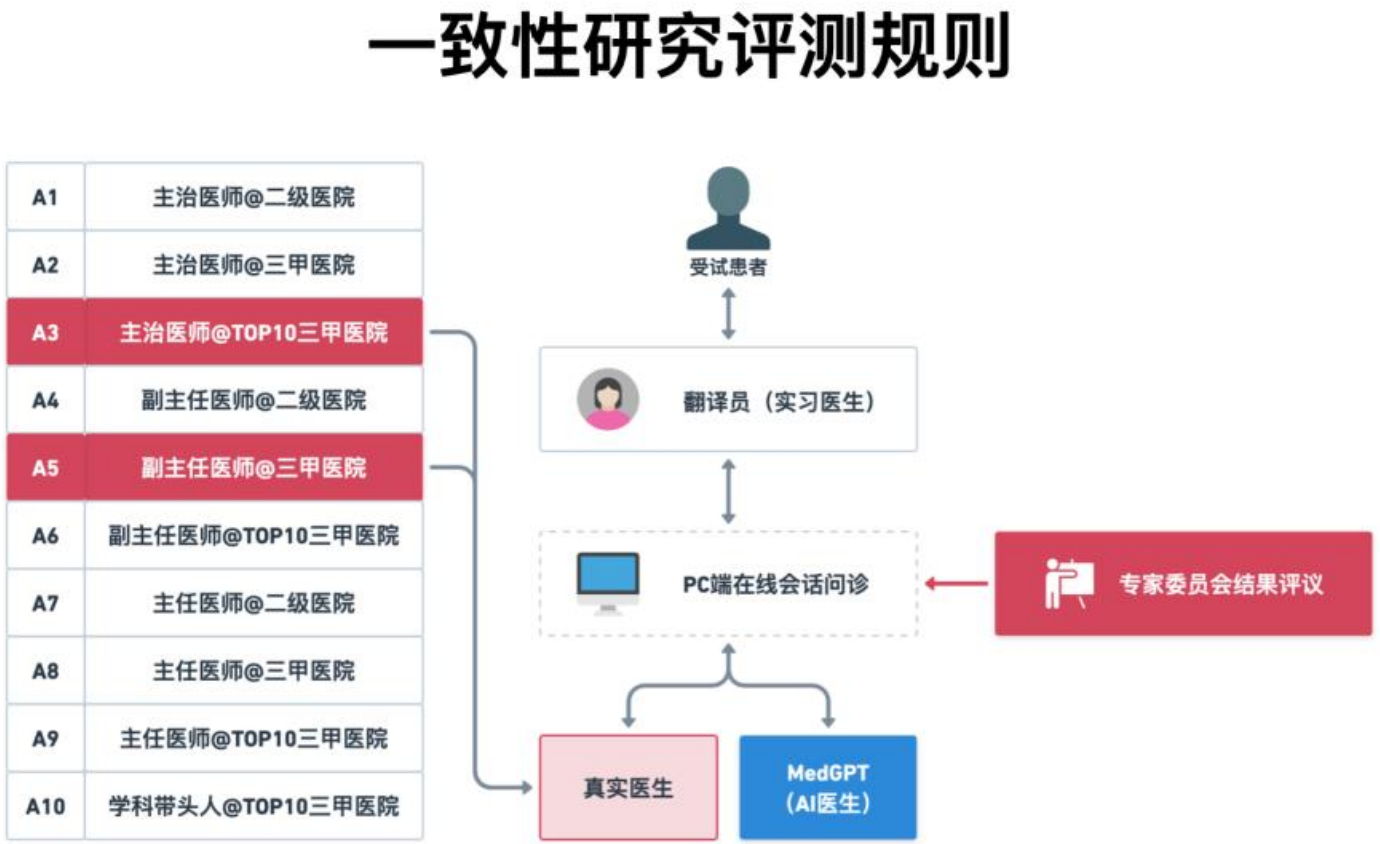
https://cmedbenchmark.llmzoo.com/static/leaderboard.html

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang*, Haizhou Li .CMB: A Comprehensive Medical Benchmark in Chinese. NAACL 2024

How to Evaluate Medical LLMs?



Expert Consistency Study with Medlinker (1)



Expert Consistency Study with Medlinker (2)



Limitations of Current Evaluation

LLMs complement doctors rather than replace them. Clinical practice is multimodal, not text-only. Diagnosis is an interactive problem under incomplete information, not a one-shot judgment. Medical knowledge is highly specialized and hard to verify. Human evaluation is costly, subjective and hard to reproduce.

Insights for Future Evaluation

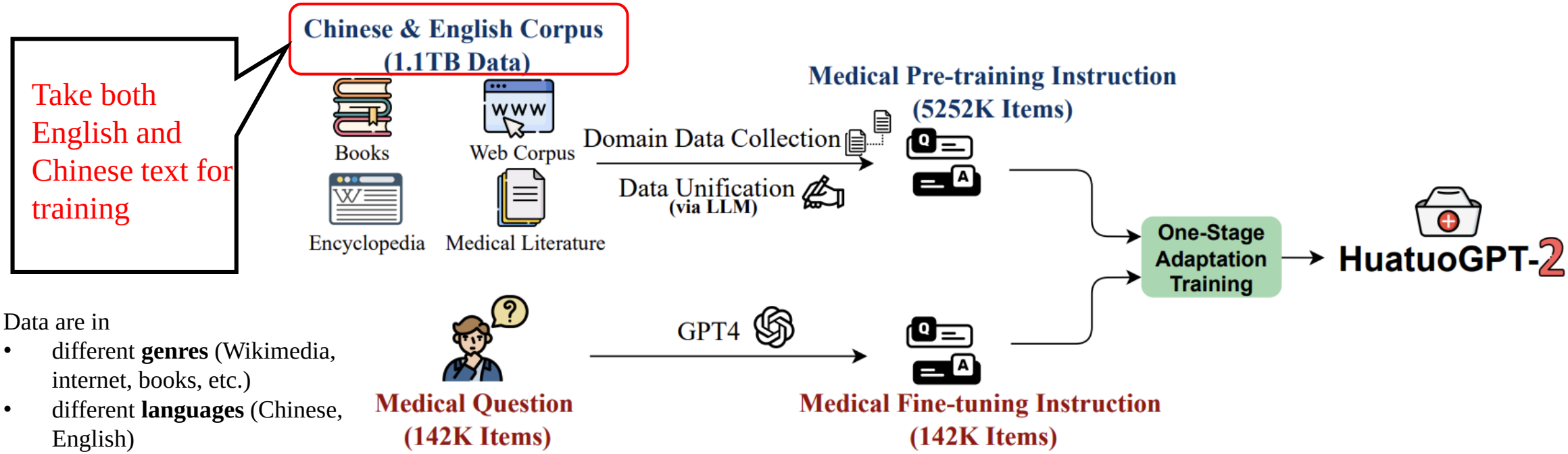
- **Difficult** enough for complex diagnosis reasoning (not saturared).
- **Multi-modal** for the input (X-ray, CT, fMRI, EEG, audio, etc)
- Evalution **without data leakage**
- Automatic (open) evaluation using (subjective) **expert's** experience [1]

[1] HealthBench: Evaluating Large Language Models Towards Improved Human Health
https://openai.com/index/healthbench/

2.3 HuatuoGPT-II: Injecting Medical Knowledge

The first HuatuoGPT talked like a doctor but failed licensing exams. Knowledge has to go into the model, not only style.

HuatuoGPT-II: Methodology



Data are in

- different **genres** (Wikimedia, internet, books, etc.)
- different **languages** (Chinese, English)
- different **format** (plain corpora, question-answer pairs, input-label pairs)

We want to unify all data in input-output pair format.

[2] Junying Chen, Xidong Wang, Anningzhe Gao#, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, **Haizhou Li, Benyou Wang#.** Huatuogpt-ii, one-stage training for medical adaption of llms. COLM 2024
https://arxiv.org/abs/2311.09774

HuatuoGPT-II: State of the Art for a While

Model	Pharmacist Licensure Examination (Pharmacy)					Pharmacist Licensure Examination (TCM)					AVG
	Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	
DISC-MedLLM	22.2	26.8	23.3	0.0	22.6	24.4	32.3	15.0	0.0	24.9	23.8
HuatuoGPT	25.6	25.5	23.3	2.6	23.4	24.1	26.8	31.6	7.5	24.9	24.2
ChatGLM2-6B	37.0	36.8	25.0	31.7	35.3	33.1	37.3	35.0	37.3	33.7	34.5
ChatGLM3-6B	39.5	39.1	10.5	0.2	34.6	31.8	38.2	25.0	20.0	32.9	33.8
Qwen-7B-chat	43.8	46.8	33.3	18.4	41.9	40.0	43.2	33.3	17.5	38.8	40.4
Qwen-14B-chat	56.2	58.6	41.7	21.1	52.7	51.3	51.0	27.5	41.7	47.9	50.3
Biachuan2-7B-Chat	51.2	50.9	30.0	2.6	44.6	48.1	46.0	35.0	7.5	42.1	43.4
Biachuan2-13B-Chat	43.8	52.7	36.7	7.9	44.2	41.3	46.4	43.3	15.0	41.7	43.0
文心一言	45.0	60.9	36.7	23.7	49.6	53.8	59.1	38.3	20.0	51.5	50.6
ChatGPT(API)	45.6	44.1	36.7	13.2	41.2	34.4	32.3	30.0	15.0	31.2	36.2
GPT-4(API)	65.1	59.6	46.7	15.8	**57.3**	40.6	42.7	33.3	17.5	38.8	48.1
HuatuoGPT-II(7B)	41.9	61.0	35.0	15.7	47.7	52.5	51.4	41.7	15.0	47.5	47.6
HuatuoGPT-II(13B)	47.5	64.1	45.0	23.7	52.9	48.8	61.8	45.0	17.5	**51.6**	**52.3**
HuatuoGPT-II(34B)	66.3	75.0	48.3	34.2	65.5	63.6	71.4	50.0	27.5	**62.5**	**64.0**

Benchmark is in Oct, where models are trained during Oct and cheating is impossible

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, **Benyou Wang**#, Haizhou Li. HuatuoGPT, towards Taming Language Model to Be a Doctor. https://arxiv.org/abs/2305.15075

Junying Chen, Xidong Wang, Anningzhe Gao#, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, **Benyou Wang#.** Huatuogpt-ii, one-stage training for medical adaption of llms. https://arxiv.org/abs/2311.09774

2024 Pharmacist Exam: Rethinking an Evolving Exam

Advanced LLMs (Over 20B Parameters)

模型	2024年中药	2024年西药	平均分
c4ai-command-r-v01	0.300	0.379	0.340
Mixtral-8x7B-Instruct-v0.1	0.325	0.354	0.340
deepseek-llm-67B-chat	0.465	0.479	0.472
GPT-4o-mini	0.469	0.648	0.559
Claude-3-5-sonnet	0.469	0.665	0.567
LLaMA-3.1-70B-Instruct	0.494	0.648	0.571
Yi-34B-Chat	0.654	0.654	0.654
GPT-4o	0.600	0.752	0.676
HuatuoGPT2-34B	0.669	0.700	**0.684**
Qwen2.5-72B-it	0.721	0.817	**0.769**

- Last year’s HuatuoGPT-2 model can still pass this year's exam, maintaining an edge over GPT-4o.
- Qwen2.5-72B scores higher, with an average accuracy of 76.8%.
- Models like LLaMA-3-70B and Claude perform moderately in Chinese, especially in Traditional Chinese Medicine.

What HuatuoGPT Taught Us

Data must look like doctors

Real doctor replies carry expertise; LLM-distilled replies carry patience and structure. Mix both.

Knowledge must be injected

Style without knowledge fails licensing exams. HuatuoGPT-II unified pre-training and fine-tuning.

Evaluation is a research problem

Benchmarks saturate and leak; doctors disagree; real diagnosis is interactive.

Deployment needs workflows

A model is not a product: triage, family doctors and records each need their own integration.

These lessons point beyond a single model — toward a generalist, trustworthy system.

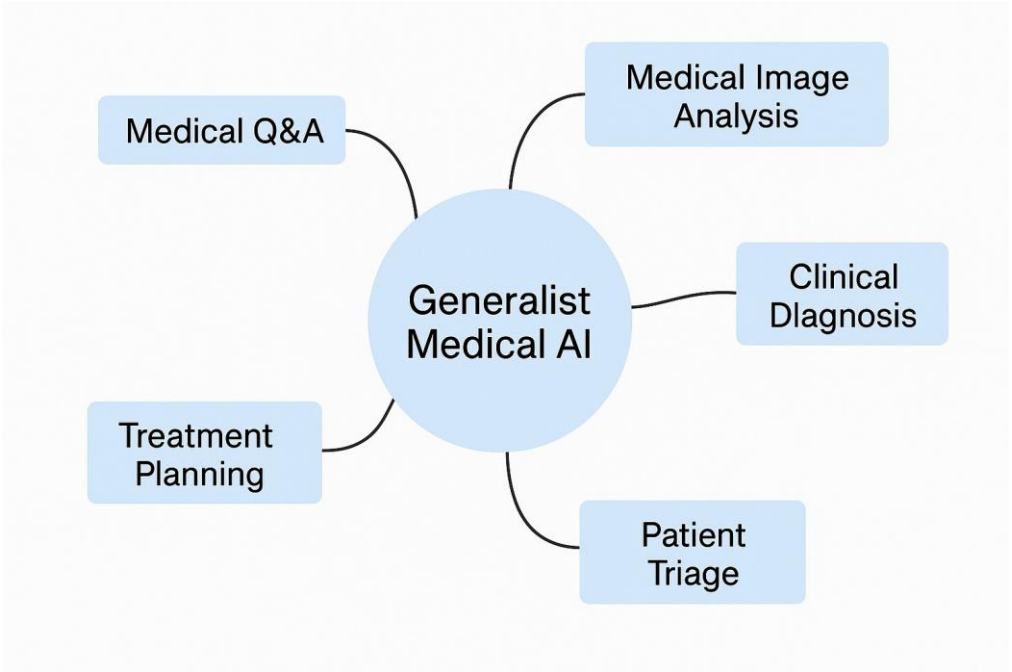
Zhang et al. HuatuoGPT (Findings of EMNLP 2023) · Chen et al. HuatuoGPT-II (COLM 2024) · Wang et al. CMB (NAACL 2024).

3 What Is Generalist Medical AI?

One model for every medical task — but the road there runs through specialists and through trust.

What Is Generalist Medical AI (GMAI)?

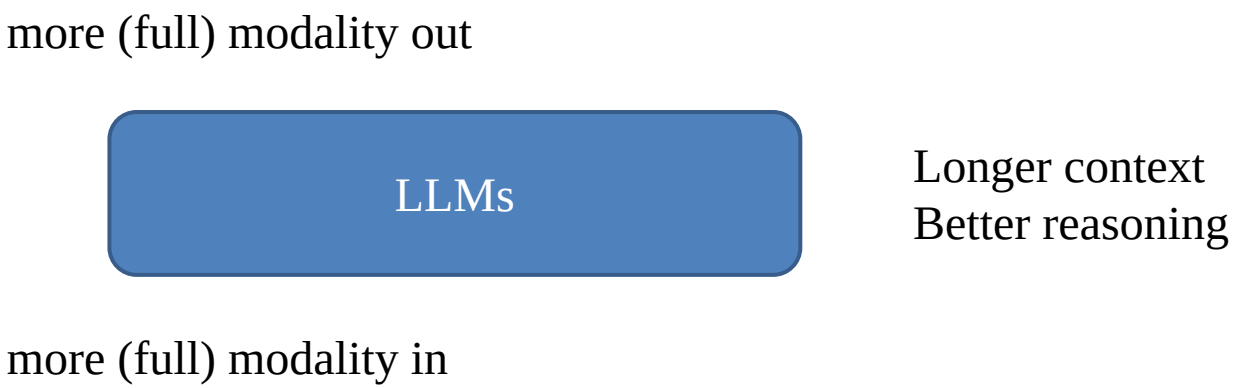
- Similarly to **A**rtificial **G**eneral **I**ntelligence (AGI)
- GMAI: one model for every medical applications with little adaptions.



- Why?
  - mutual benefits between tasks (especially when data are limited)
  - deployment efficiency

Why Is Generalist Medical AI Language-Centric?

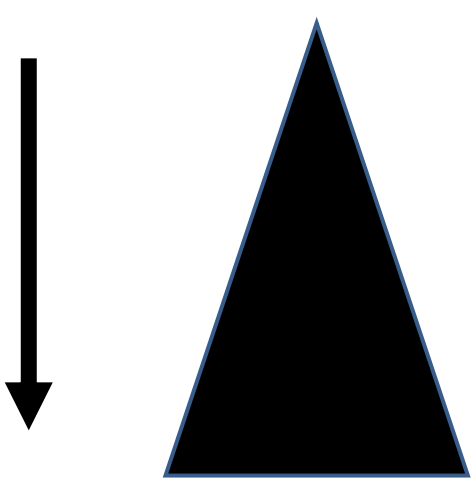
- Language has massive **knowledge**
  - "*The sun rises in the east*"
- Language is a great tool for **communication**
  - The interface between Humans and Machines
- Language can **reason and plan**
  - Language is the medium of thought



GMAI: Top-Down or Bottom-Up?

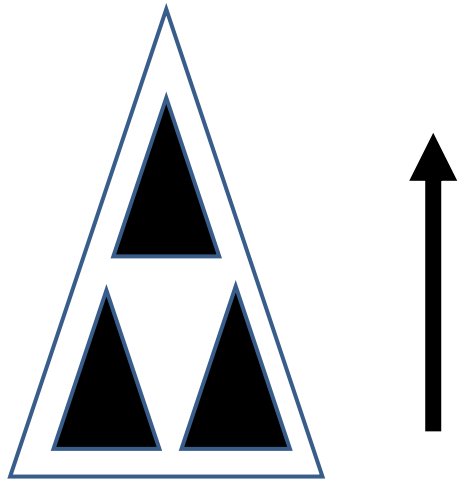
- GMAI is the final goal
- The road towards this goal might involves different specialized Multi-modal LLMs

Open AI develop one model for all



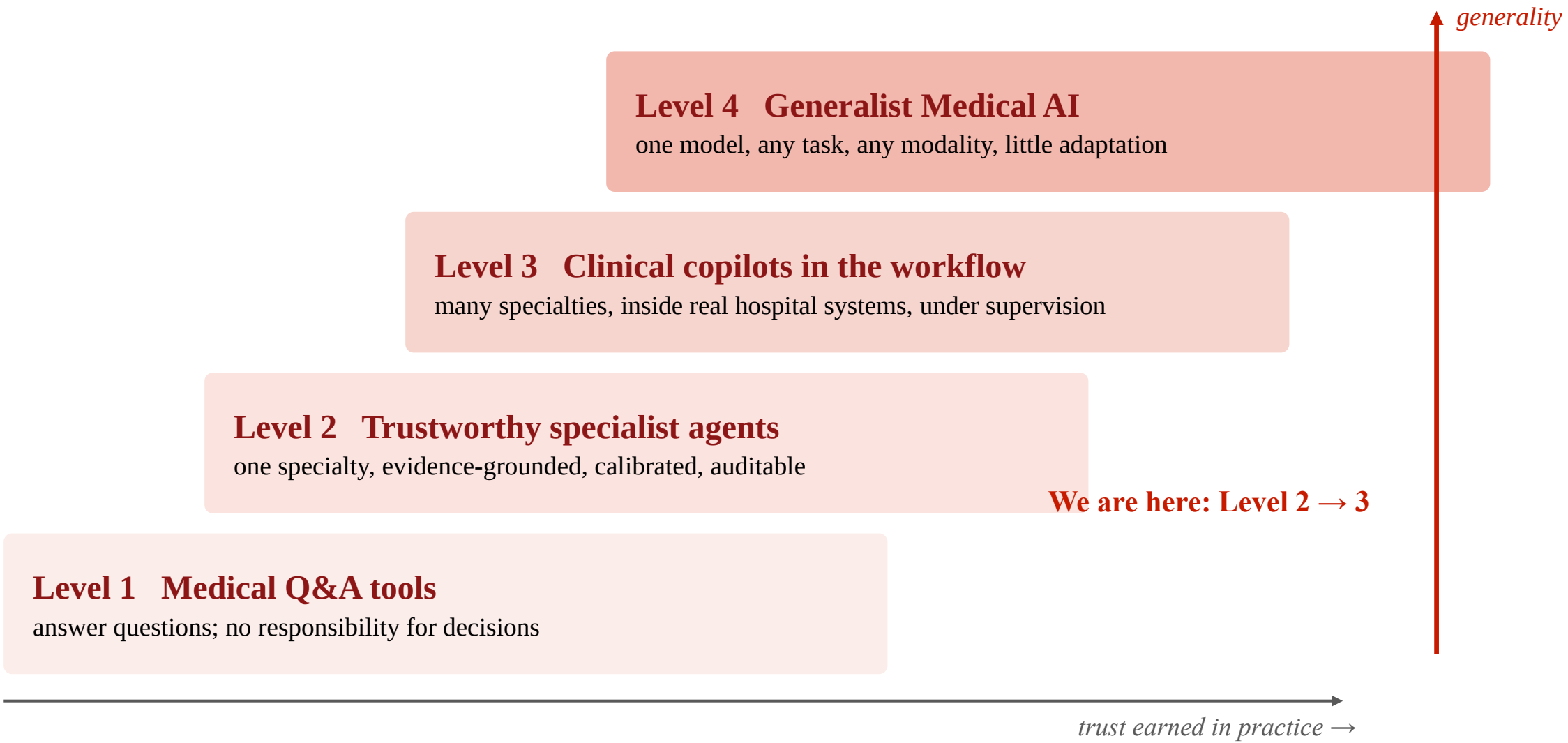
Better model backbone

Better insights/data/benchmark



Different specialized models and merge them

A Ladder towards Generalist Medical AI



Roadmap to Generalist Medical AI

- **What is Generalist Medical AI?**
- Roadmap to Generalist Medical AI
  - **Perception** :
    - I) Multi-modal Perception
    - II) Longer Context
  - **Cognition**: III) Better Reasoning
  - **Interaction**:
    - IV) Multi-modal Generation
    - V) Agent Simulation : Twin Hospital
  - **Multilingual Support**
    - VI) Multilingual Medical LLMs: Apollo
- Challenges of GMAI and Solutions
  - I) Generalist Medical LLMs vs. Specialized Medical LLMs
  - II) Trustworthiness and interpretation
  - III) Lack Challenging Benchmark

4 The Roadmap: Perception, Reasoning, Generation, Languages

A generalist doctor must see, hear, reason, explain and speak every patient's language.

In this part

4.1 Perception and Long Context

4.2 Reasoning

4.3 Generation and Simulation

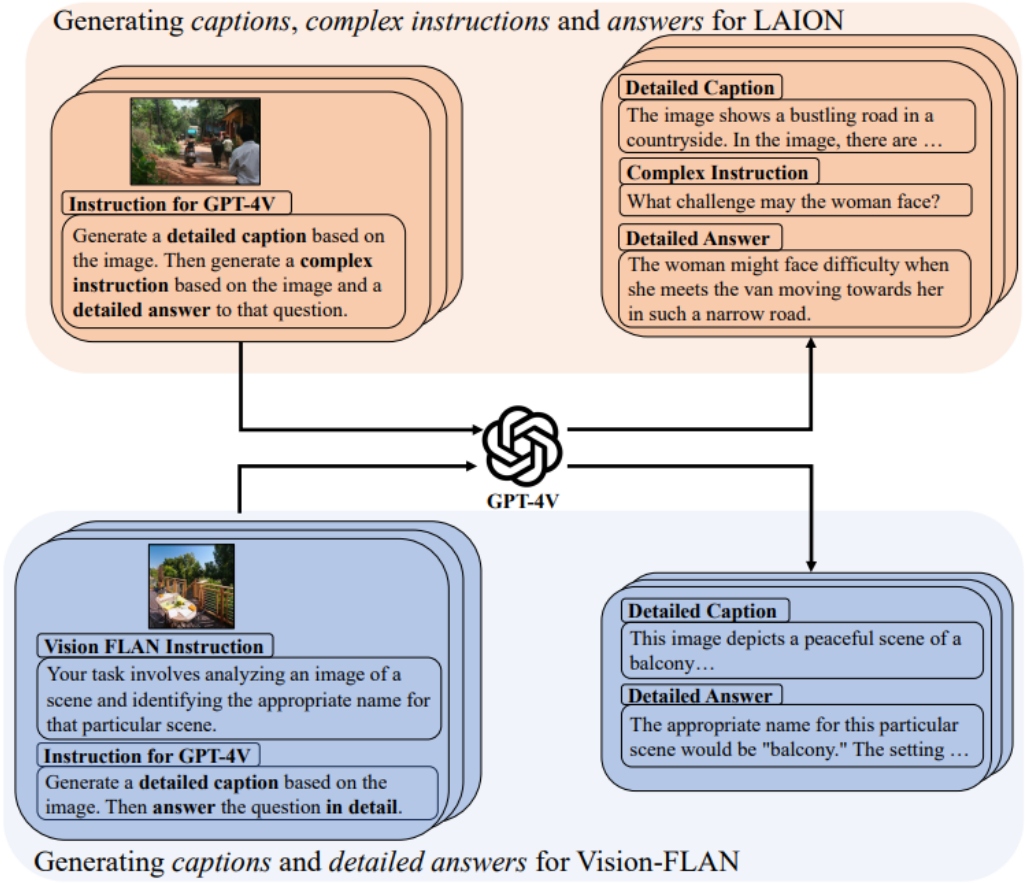
4.4 Multilinguality

Perception: Multimodal LLMs

- High-quality image–text data and models (ALLaVA)
- Long-context multimodal evaluation (MileBench, COLM 2024)
- Motion understanding (MotionLLM)
- MLLM evaluation (MLLM-Bench)
- Feedback learning for MLLMs (VLFeedback)
- Speech understanding (Soundwave)
- End-to-end speech LLM (EchoX)
- LLM-based image generation (Janus-4o)
- Medical video generation (MedGen)
- Full-modality TCM model (ShizhenGPT)
- Video reasoning (Video-R1)

[1] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, **Benyou Wang**. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684
[2] D Song, S Chen, GH Chen, F Yu, X Wan**, B Wang**. Milebench: Benchmarking mllms in long context, arXiv preprint arXiv:2404.18532
[3] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, Lei Zhang. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. https://arxiv.org/abs/2405.20340
[4] Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang. Mllm-bench, evaluating multi-modal llms using gpt-4v. https://arxiv.org/abs/2311.13951
[5] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong. Silkie: Preference distillation for large visual language models. https://arxiv.org/abs/2312.10665
[6] Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, Haizhou Li. Soundwave: Less is More for Speech-Text Alignment in LLMs. https://arxiv.org/abs/2502.12900
[7] Yuhao Zhang, Yuhao Du, Zhanchen Dai, Xiangnan Ma, Kaiqi Kou, Benyou Wang, Haizhou Li. EchoX: Towards Mitigating Acoustic-Semantic Gap via Echo Training for Speech-to-Speech LLMs. https://www.arxiv.org/abs/2509.09174
[8] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, Benyou Wang. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation. https://arxiv.org/abs/2506.18095
[9] Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, Benyou Wang. MedGen: Unlocking Medical Video Generation by Scaling Granularly-annotated Medical Videos. https://arxiv.org/abs/2507.05675
[10] Junying Chen, Zhenyang Cai, Zhiheng Liu, Yunjin Yang, Rongsheng Wang, Qingying Xiao, Xiangyi Feng, Zhan Su, Jing Guo, Xiang Wan, Guangjun Yu, Haizhou Li, Benyou Wang. ShizhenGPT: Towards Multimodal LLMs for Traditional Chinese Medicine. https://arxiv.org/abs/2508.14706
[11] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. https://arxiv.org/pdf/2503.21776

Perception: Starting from Data (ALLaVA-4V)



Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, Benyou Wang. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684

Perception: Multimodal LLMs Trained on ALLaVA

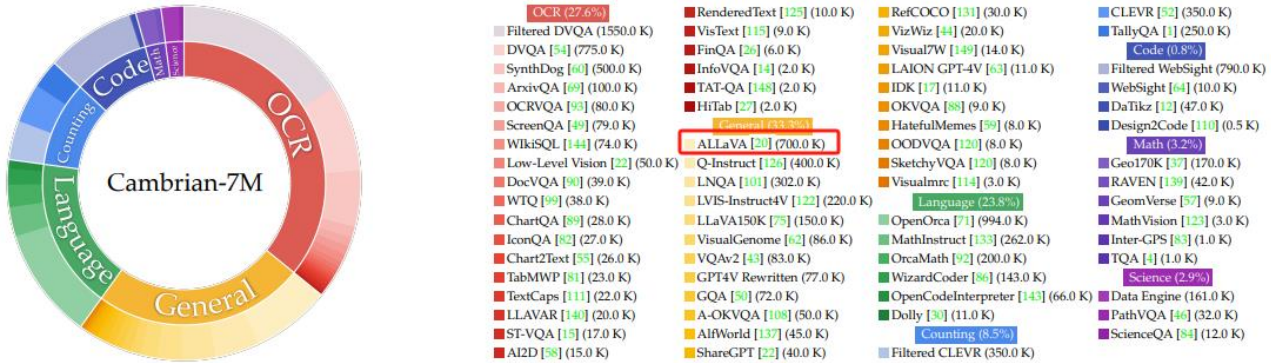


Figure 9 | **Cambrian-7M: A Large-Scale Curated Instruction Tuning Dataset for MLLM. Left:** The inner circle shows the original distribution of Cambrian-10M. The outer circle shows the curated Cambrian-7M. **Right:** All the data sources in the Cambrian dataset as well as the ones filtered in data curation.

***Cambrian*-1: A Fully Open, *Vision-Centric* Exploration of Multimodal LLMs**

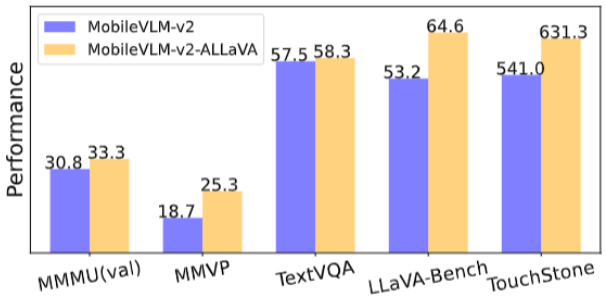
Shengbang Tong*, Ellis Brown*, Penghao Wu*, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, Saining Xie†

New York University

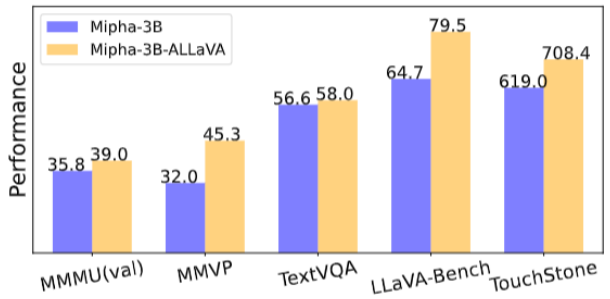
Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, **Benyou Wang**. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684

Better Data, Better Models



(a) MobileVLM-v2

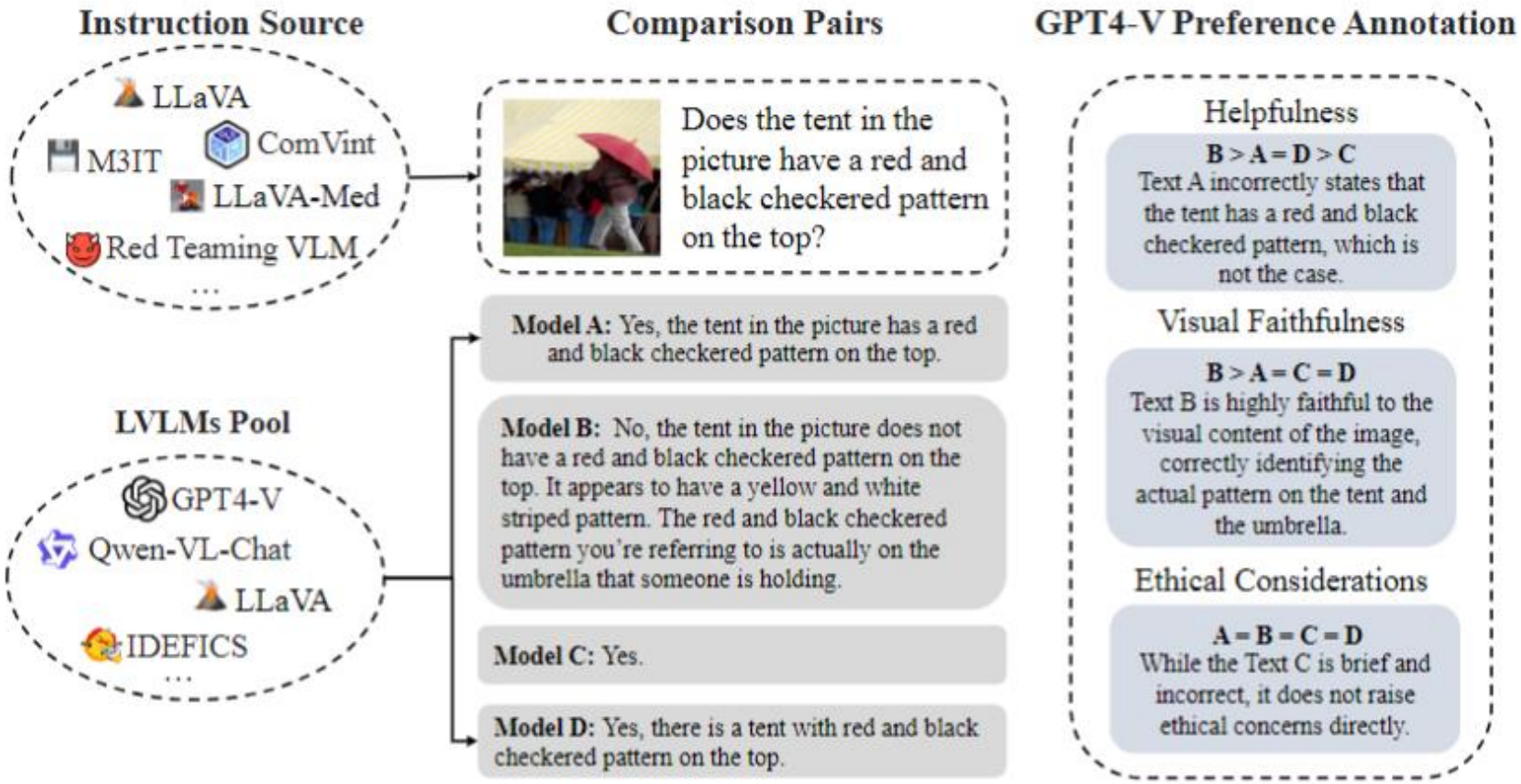


(b) Mipha-3B

Model	LM Backbone	Benchmarks											
		Text	Multimodal (Close-ended)								Multimodal (Open-ended)		
		Vicuna-80	MMB	$SEED^{v1}_{img}$	MM-Vet	$MMMU^{val}$	MME	VQA^{T}	GQA	EMT^{c10}	MB	TS	$LLaVA^{W}$
InstructBLIP	Vicuna-13B	-	44.0	-	25.6	-	1212.8	50.7	49.5	-	4.0	552.4	58.2
BLIP-2-T5-XL	FLAN-T5-XL(4B)	-	-	49.7	22.4	34.4	-	-	-	-	2.1	-	-
Qwen-VL-Chat	Qwen-7B	-	60.6	65.4	-	35.9	1487.5	61.5	57.5	-	6.2	711.6	-
LLaVA-v1.7B	Vicuna-7B	-	64.3	-	31.1	-	1510.7	58.2	62.0	-	-		65.4
LLaVA-v1.5 13B	Vicuna-13B	22.50	67.7	68.2	35.4	36.4	1531.3	61.3	63.3	85.0	7.4	637.7	70.7
LVIS-Inst4V 7B	Vicuna-7B	-	66.2	-	31.5	-	1528.2	58.7	62.6	-	6.0	-	67.0
LVIS-Inst4V 13B	Vicuna-13B	-	68.0	-	37.4	-	1574.9	62.5	63.6	-	-	-	71.3
ShareGPT4V 7B	Vicuna-7B	-	68.8	69.7	37.6	-	1943.8	60.4	63.3	-	-	-	72.6
ShareGPT4V 13B	Vicuna-13B	-	71.2	70.8	43.1	-	1921.9	62.2	64.8	-	-	-	79.9
TinyGPT-V	Phi2-2.7B	-	-	-	-	-	-	-	33.6	-	-	-	-
MobileVLM	MobileLLaMA-2.7B	-	59.6	-	-	-	1288.9	47.5	-	-	-	-	-
LLaVA-Phi	Phi2-2.7B	-	59.8	-	28.9	-	1335.1	48.6	-	-	-	-	-
ALLaVA	Phi2-2.7B	48.8	64.0	65.2	32.2	**35.3**	**1623.2**	49.5	48.8	**90.2**	6.7	632.0	69.4
ALLaVA-Longer	Phi2-2.7B	**52.5**	**64.6**	**65.6**	**35.5**	33.2	1564.6	**50.3**	**50.0**	85.9	**8.8**	**636.5**	**71.7**

It largely improve performance by replacing original data with ours.

VLFeedback: Preference Data for Vision–Language Models

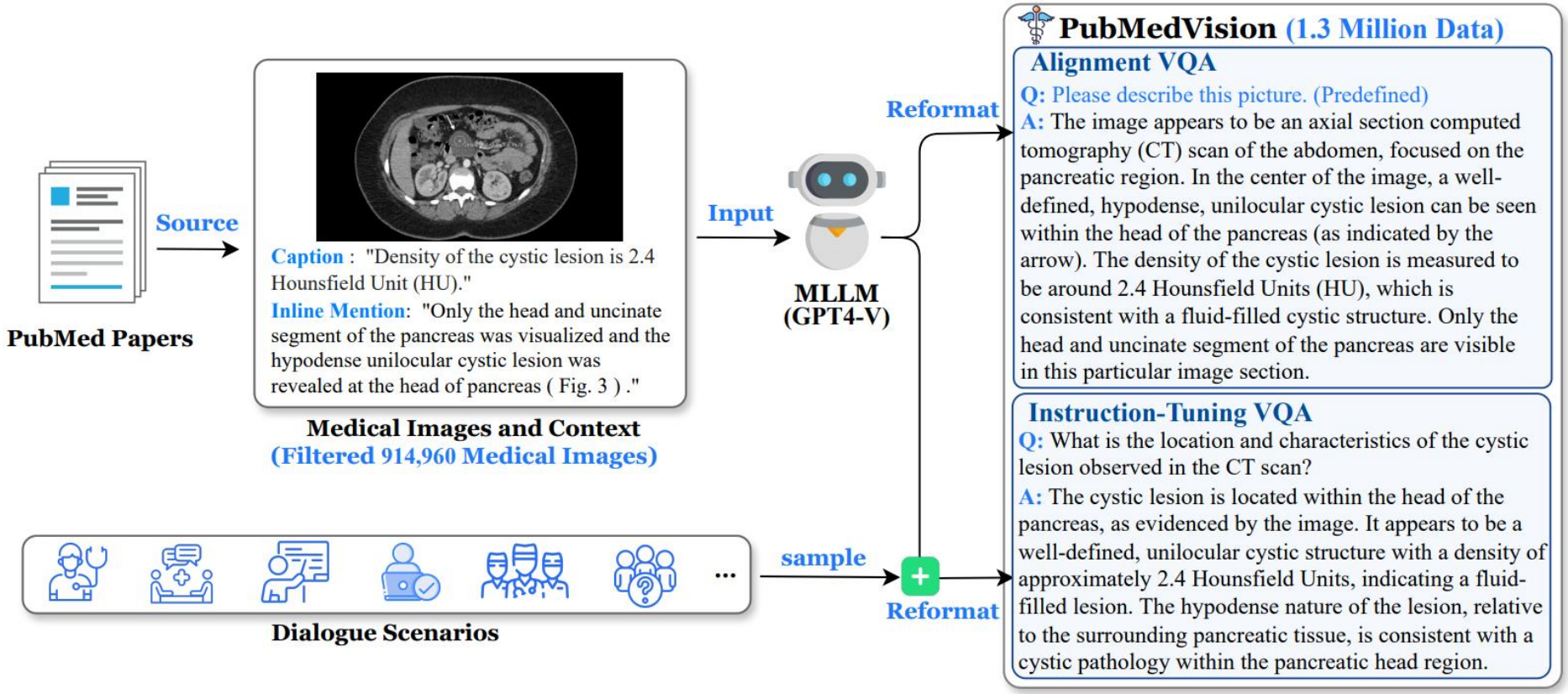


Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, Qi Liu. Silkie: Preference distillation for large visual language models. https://arxiv.org/abs/2312.10665

DPO with VLFeedback on Qwen-VL

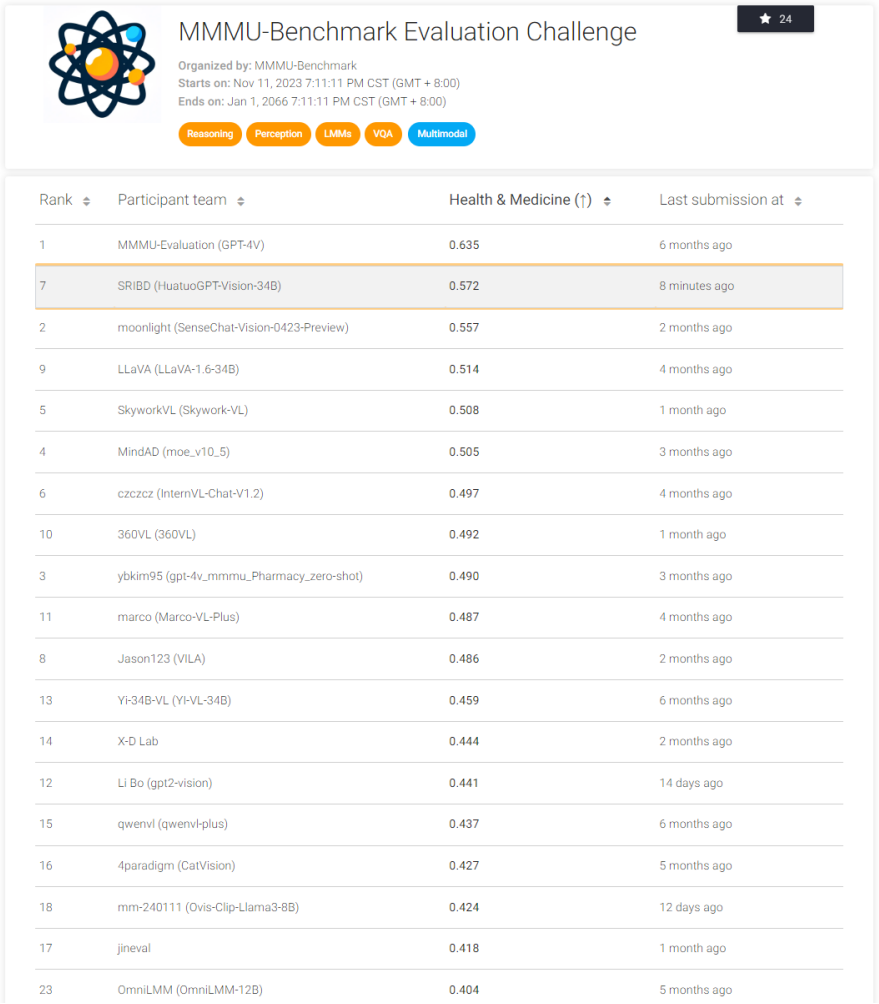
Model	MME^P	MME^C	MMHal-Bench	MM-Vet	MathVista	MMMU
LLaVA-SFT*	1315.7	260.0	1.76	29.4	25.2	33.1
+ LLaVA-RLHF*	1203.3 (↓)	273.2 (↑)	2.05 (↑)	29.0 (↓)	25.0 (↓)	30.6 (↓)
LLaVA-v1.5*	1510.7	316.1	2.42	30.5	26.7	35.3
+ POVID*	1423.9 (↓)	334.6 (↑)	2.69 (↑)	31.8 (↑)	26.1 (↓)	34.0 (↓)
+ HA-DPO*	1502.6 (↓)	313.9 (↓)	2.24 (↓)	29.4 (↓)	26.6 (↓)	34.9 (↓)
Qwen-VL-Chat	1439.1	362.5	<u>2.89</u>	45.7	<u>40.0</u>	<u>35.9</u>
+ SFT (ShareGPT4V)*	1527.4 (↑)	-	-	45.9 (↑)	-	-
+ DPO (Longest as Best)	1333.5 (↓)	343.6 (↓)	2.73 (↓)	<u>46.8</u> (↑)	37.4 (↓)	34.2 (↓)
+ DPO (GPT-4V as Best)	1210.0 (↓)	248.6 (↓)	2.76 (↓)	45.9 (-)	37.7 (↓)	32.8 (↓)
Silkie (Ours)	**1539.6** (↑)	**397.1** (↑)	**3.02** (↑)	**49.9** (↑)	**42.5** (↑)	**37.4** (↑)

HuatuoGPT-Vision: Injecting Medical Visual Knowledge



Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, Benyou Wang. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. EMNLP 2024 https://arxiv.org/abs/2406.19280

Impact of HuatuoGPT-Vision



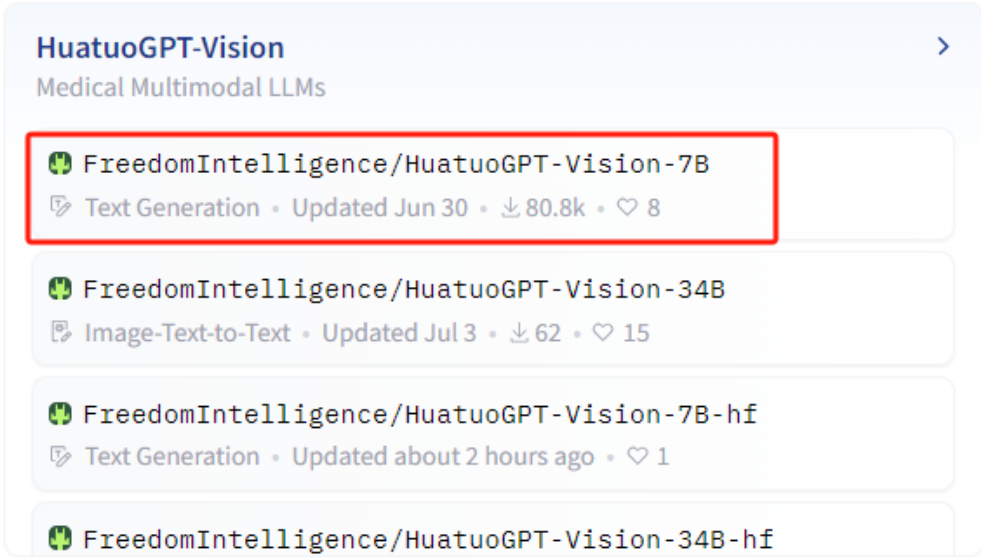
MMMU-Benchmark Evaluation Challenge

★ 24

Organized by: MMMU-Benchmark
Starts on: Nov 11, 2023 7:11:11 PM CST (GMT + 8:00)
Ends on: Jan 1, 2066 7:11:11 PM CST (GMT + 8:00)

Reasoning | Perception | LMMs | VQA | Multimodal

Rank	Participant team	Health & Medicine (↑)	Last submission at
1	MMMU-Evaluation (GPT-4V)	0.635	6 months ago
7	SRIBD (HuatuoGPT-Vision-34B)	0.572	8 minutes ago
2	moonlight (SenseChat-Vision-0423-Preview)	0.557	2 months ago
9	LLaVA (LLaVA-1.6-34B)	0.514	4 months ago
5	SkyworkVL (Skywork-VL)	0.508	1 month ago
4	MindAD (moe_v10_5)	0.505	3 months ago
6	czczcz (InternVL-Chat-V1.2)	0.497	4 months ago
10	360VL (360VL)	0.492	1 month ago
3	ybkim95 (gpt-4v_mmmu_Pharmacy_zero-shot)	0.490	3 months ago
11	marco (Marco-VL-Plus)	0.487	4 months ago
8	Jason123 (VILA)	0.486	2 months ago
13	Yi-34B-VL (YI-VL-34B)	0.459	6 months ago
14	X-D Lab	0.444	2 months ago
12	Li Bo (gpt2-vision)	0.441	14 days ago
15	qwenvl (qwenvl-plus)	0.437	6 months ago
16	4paradigm (CatVision)	0.427	5 months ago
18	mm-240111 (Ovis-Clip-Llama3-8B)	0.424	12 days ago
17	jineval	0.418	1 month ago
23	OmniLMM (OmniLMM-12B)	0.404	5 months ago



HuatuoGPT-Vision
Medical Multimodal LLMs

FreedomIntelligence/HuatuoGPT-Vision-7B
Text Generation • Updated Jun 30 • ↓ 80.8k • ♡ 8

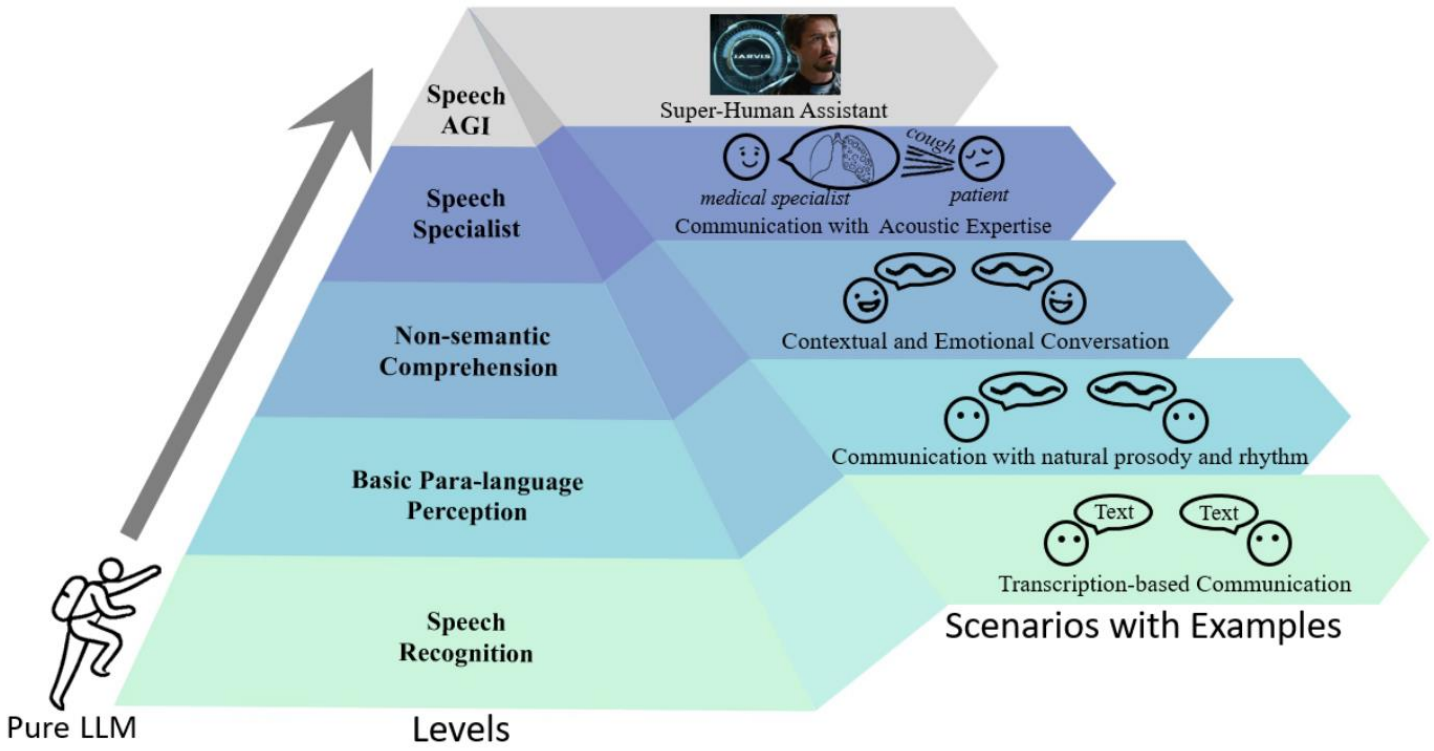
FreedomIntelligence/HuatuoGPT-Vision-34B
Image-Text-to-Text • Updated Jul 3 • ↓ 62 • ♡ 15

FreedomIntelligence/HuatuoGPT-Vision-7B-hf
Text Generation • Updated about 2 hours ago • ♡ 1

FreedomIntelligence/HuatuoGPT-Vision-34B-hf

Maximum **200K model downloading** per month

Perception: Speech (1) Five Levels of Speech Understanding



Five Levels for Speech Understanding using LLMs

Fan Bu, Yuhao Zhang, Xidong Wang, **Benyou Wang**, Qun Liu, Haizhou Li. Roadmap towards Superhuman Speech Understanding using Large Language Models. https://arxiv.org/abs/2410.13268

Perception: Speech (2) Soundwave

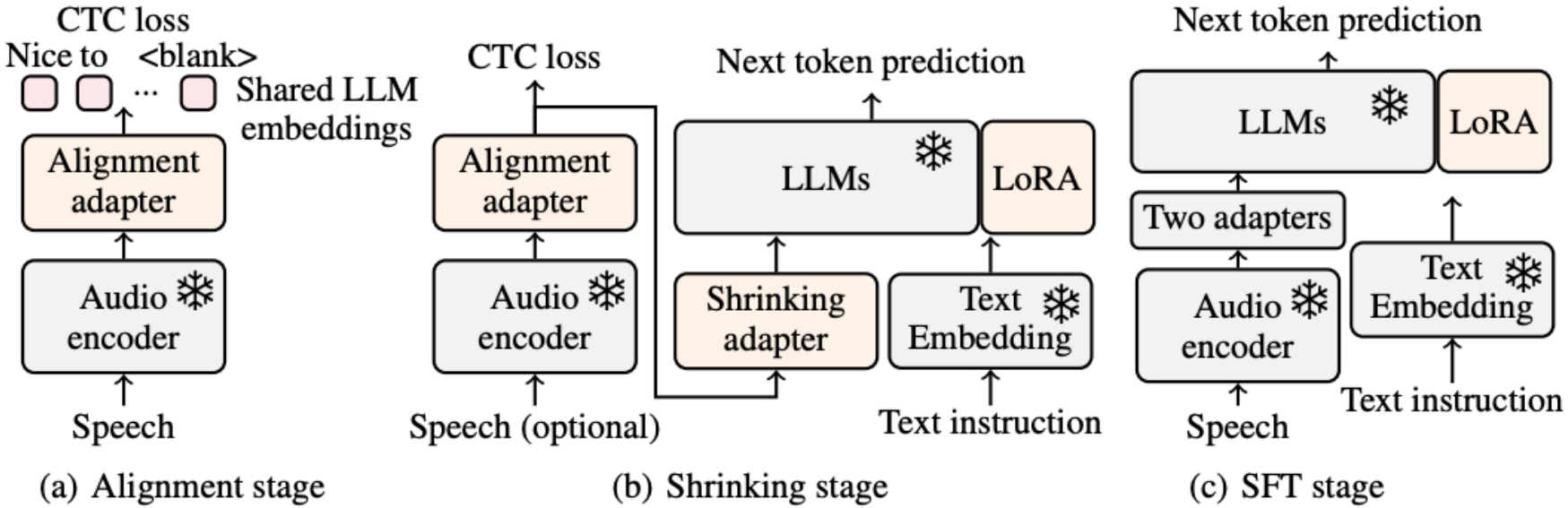
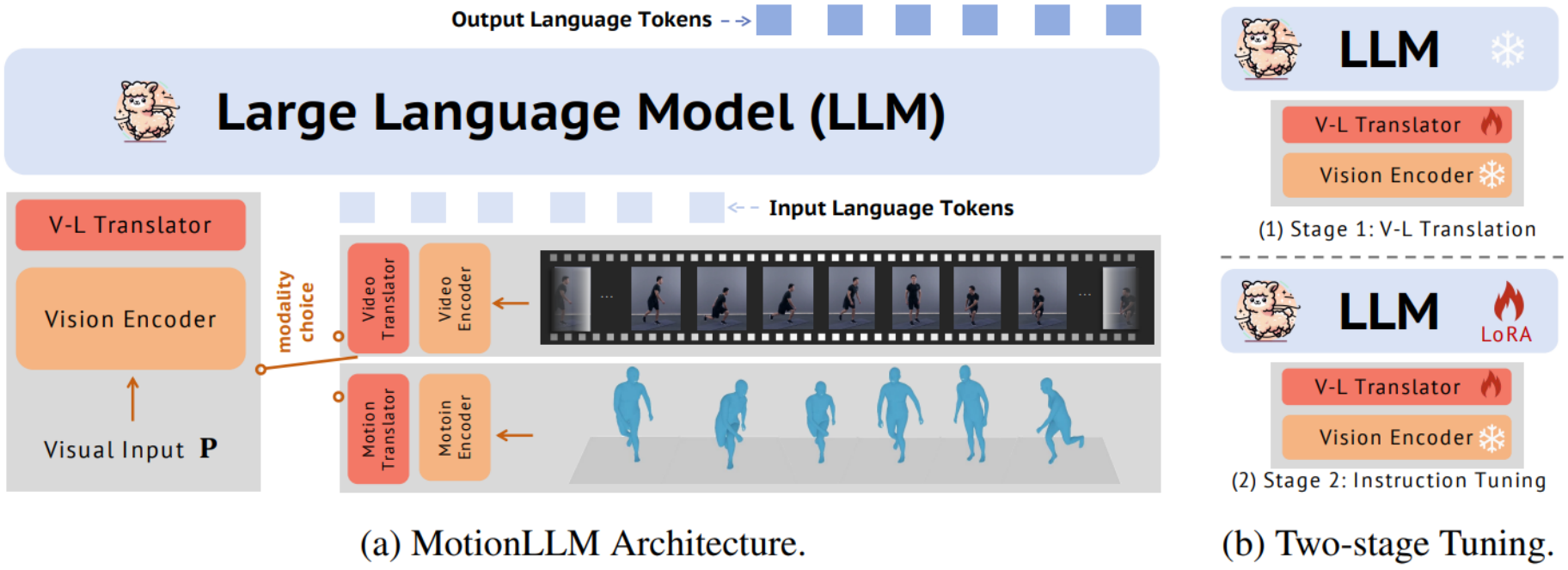


Figure 2: Training progress of Soundwave. The gray modules are frozen while the orange modules are updated.

Achieving better performance **using 1/50 data** compared to Qwen 2- Audio

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, **Benyou Wang**, Haizhou Li. Soundwave: Less is More for Speech-Text Alignment in LLMs. ACL 2025

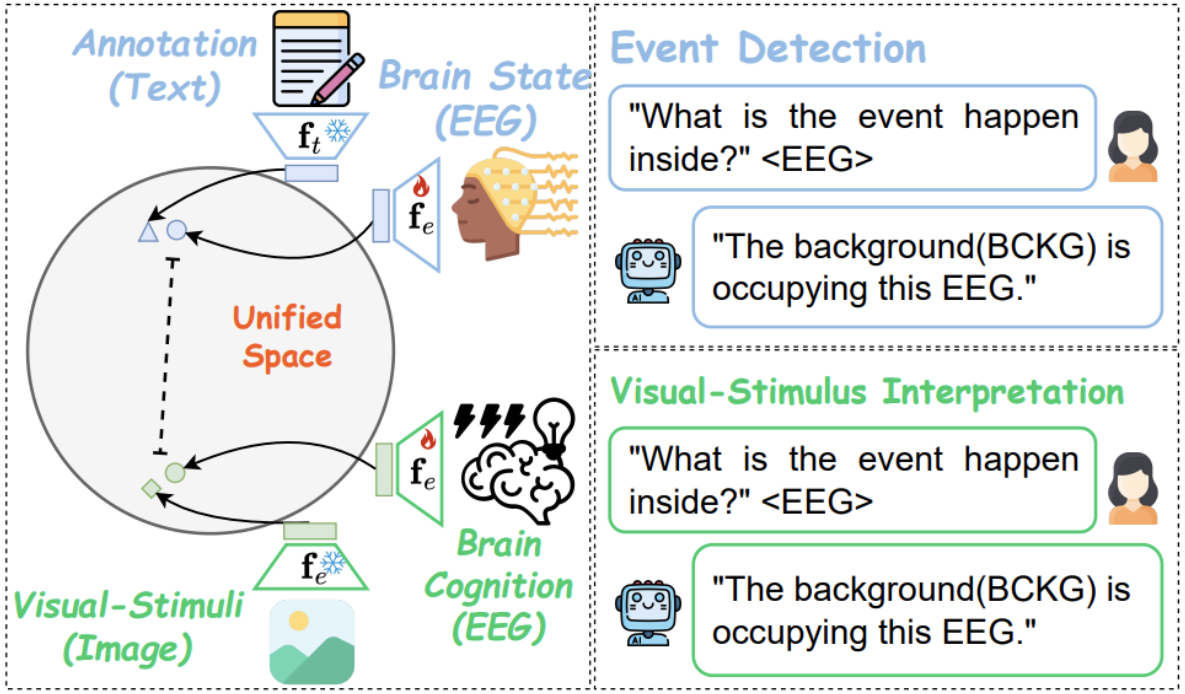
Perception: Motion (MotionLLM)



(a) MotionLLM Architecture. (b) Two-stage Tuning.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, Lei Zhang.
MotionLLM: Understanding Human Behaviors from Human Motions and Videos. TMAPI

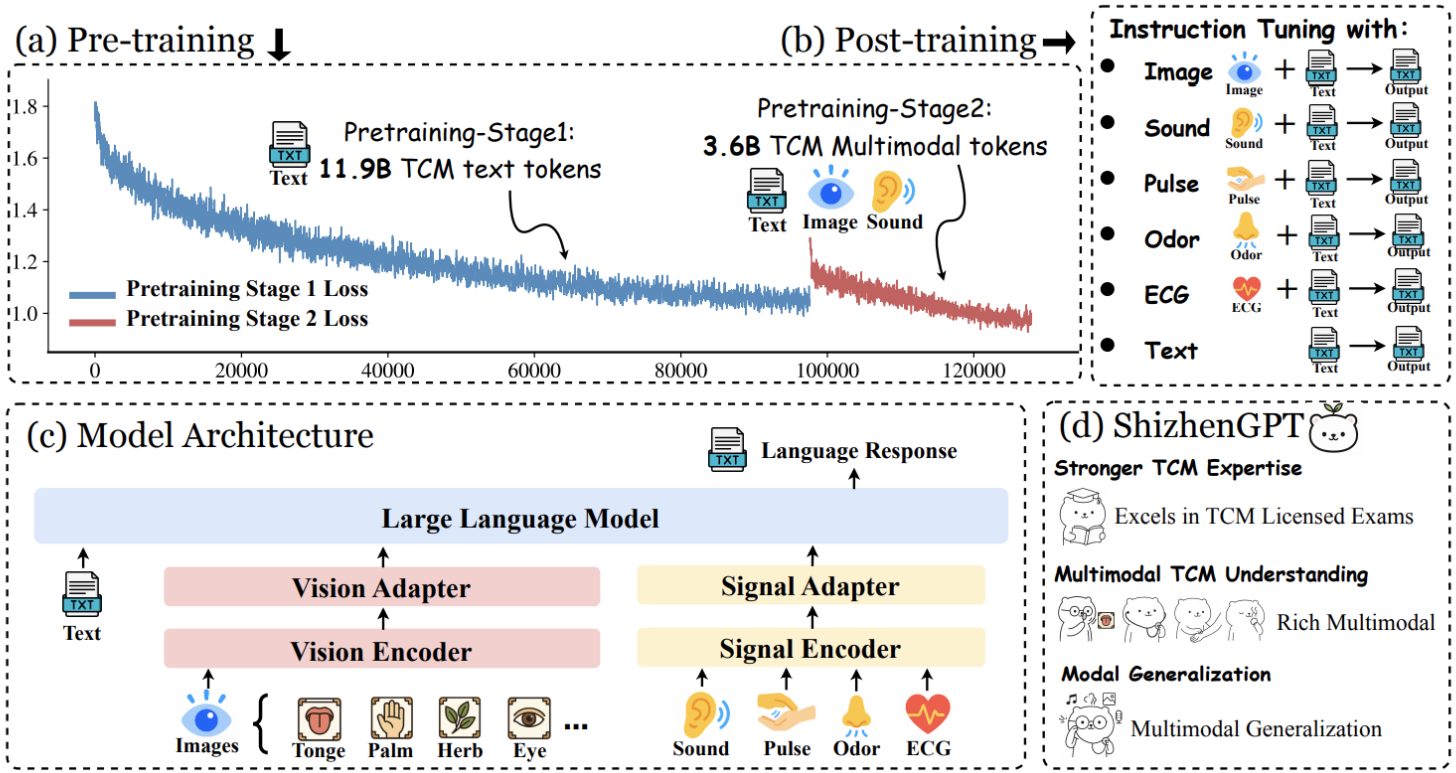
Perception: EEG (WaveMind)



Alingning EEG with texts and images!

Ziyi Zeng, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Yipeng Liu, Siqi Cai, **Benyou Wang**, Zhiguo Zhang, Haizhou Li. WaveMind: Towards a Conversational EEG Foundation Model Aligned to Textual and Visual Modalities. https://arxiv.org/abs/2510.00032

Perception: Traditional Chinese Medicine — ShizhenGPT

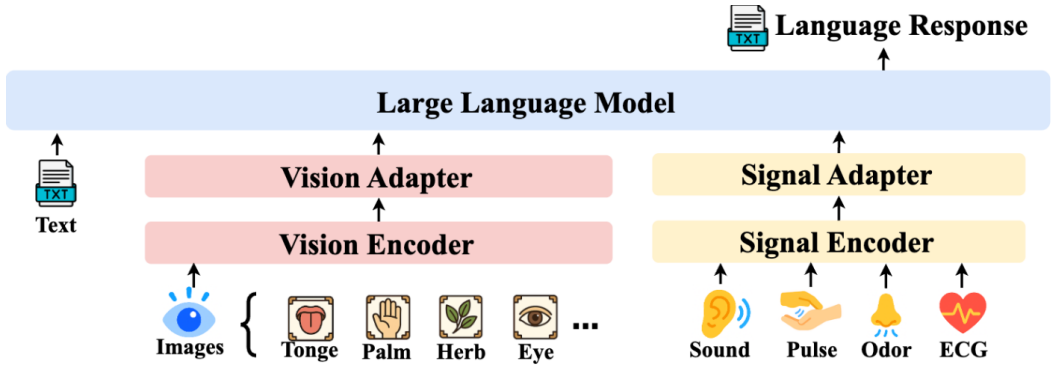


First unified model that supports "**seeing, smelling, listening, and pulse**" for Traditional Chinese Medicine

Junying Chen, Zhenyang Cai, Zhiheng Liu, Yunjin Yang, Rongsheng Wang, Qingying Xiao, Xiangyi Feng, Zhan Su, Jing Guo, Xiang Wan, Guangjun Yu, Haizhou Li, **Benyou Wang**. ShizhenGPT: Towards Multimodal LLMs for Traditional Chinese Medicine. https://arxiv.org/abs/2508.14706

ShizhenGPT: Architecture

ShizhenGPT: Architecture



The model has three core modules:

1. **LLM backbone: understands, integrates and reasons over medical knowledge.**
2. **Vision encoder: turns images and video into semantic vectors — “looking” (inspection).**
3. **Signal encoder: one encoder for sound, odor, pulse waves and ECG — “listening/smelling” and “pulse taking”.**

ShizhenGPT: Understanding Physiological Signals

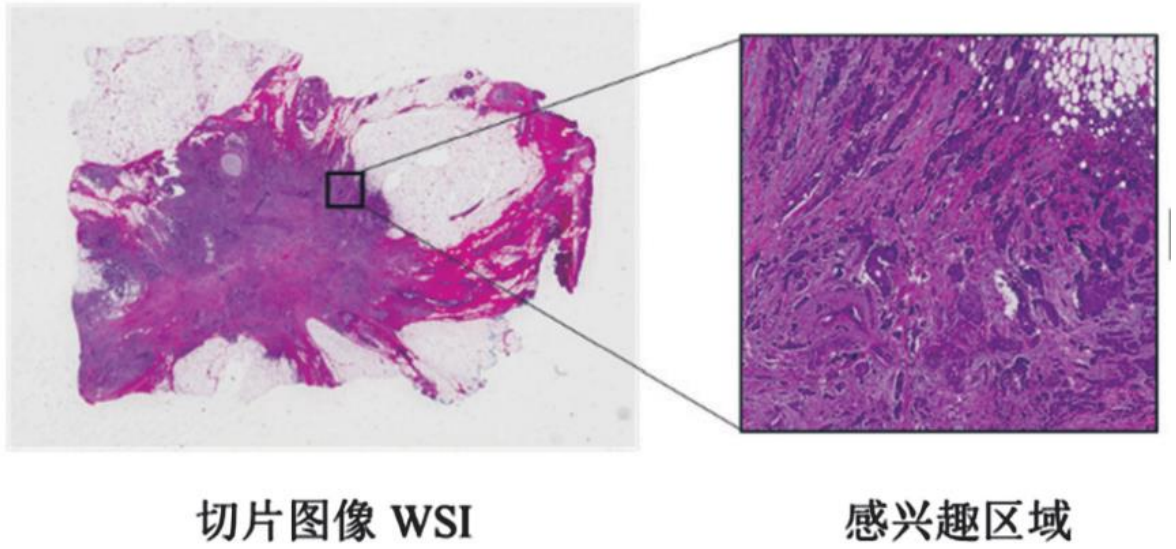
ShizhenGPT: Understanding Physiological Signals

Modality	Task	Classes	Random Baseline	TCM-Omni
Pulse	Pregnancy Detection via Pulse Diagnosis	2	50.0	80.5
	Disease Classification via Pulse Diagnosis	9	11.1	20.6
Odor	Disease Classification via Odor Analysis	7	14.3	48.8
Sound	COVID-19 Detection from Cough Sound	2	50.0	58.7
	Cardiac Abnormality Detection from Heart Sound	2	50.0	62.9
ECG	ECG Abnormality Detection	2	50.0	71.5
	ECG Heart Beat Classification	5	20.0	83.1

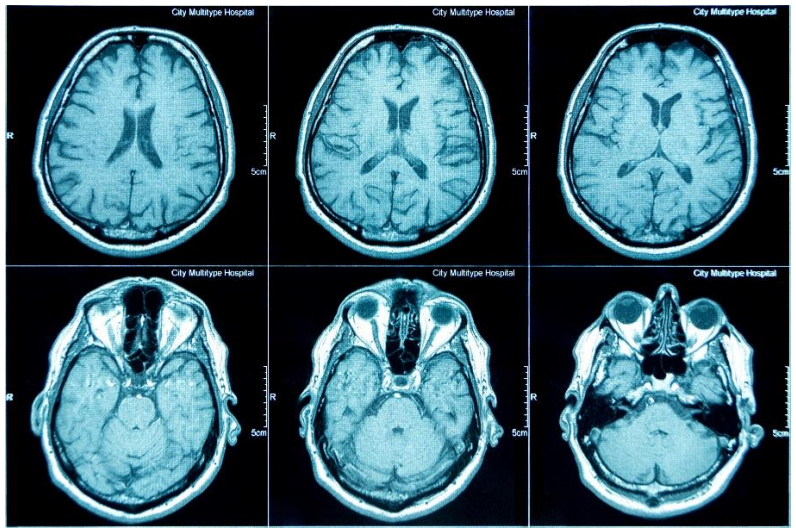
On sound, odor, pulse and ECG tasks the model shows strong perception and TCM inference — e.g., 80% accuracy detecting pregnancy from the pulse.

Long Context: Medical Images Are Huge

Motivations:



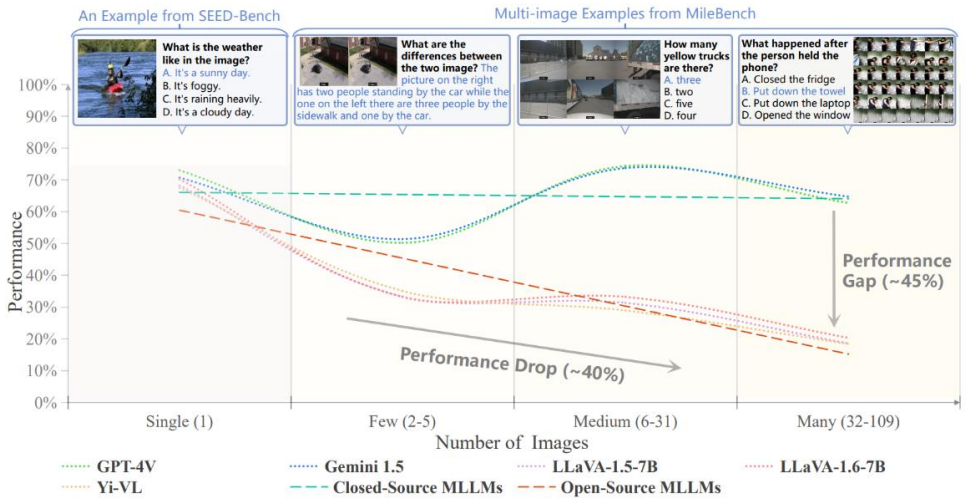
Pathology images are super **large**

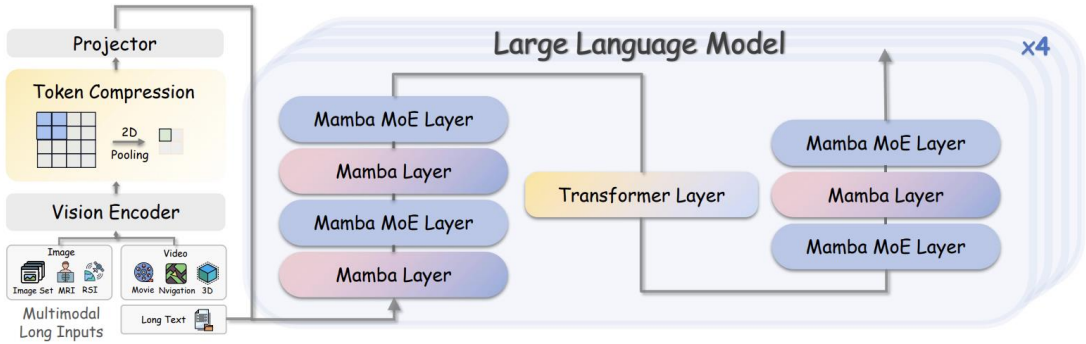


It sometimes involves **multiple** images

This require large language models to receive longer input (which are quadratic to sequence length, **challenging**!!!)

Long Context: LongLLaVA





Open-source models process long context **poorly**!

Our models (**LongLLaVA**) did it **effectively** and **efficiently**

Dingjie Song, et.al **Benyou Wang**. MILEBENCH: Benchmarking MLLMs in Long Context, COLM 2024. https://arxiv.org/abs/2404.18532

Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, **Benyou Wang**. LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via a Hybrid Architecture. Findings of EMNLP 2025

Long Context Application 1: High-Resolution Images

- Nested patching (requires longer context)

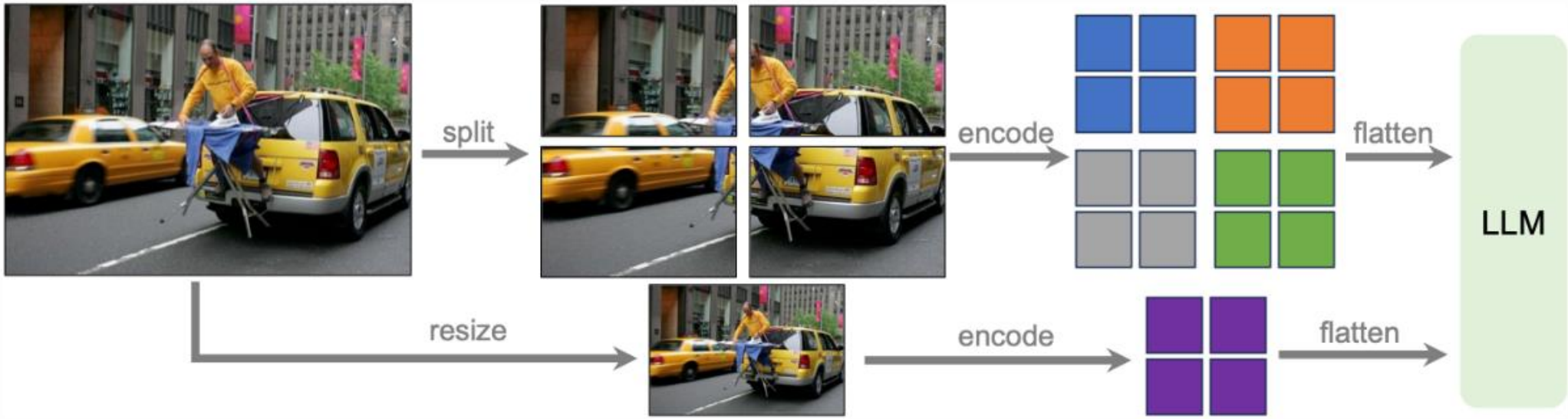
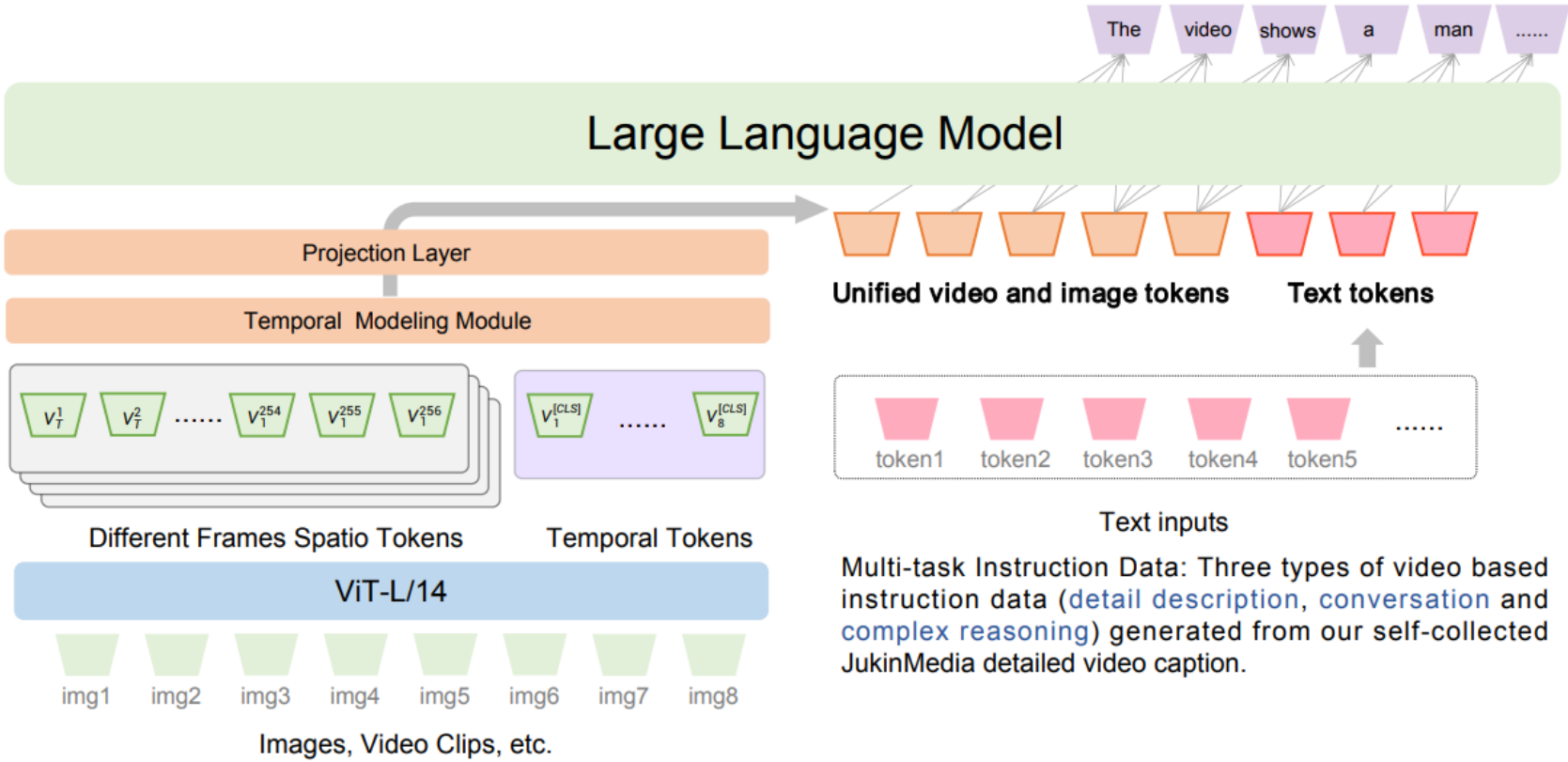


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

Finer patching: 1×1, 2×2, 4×4, 8×8, 16×16

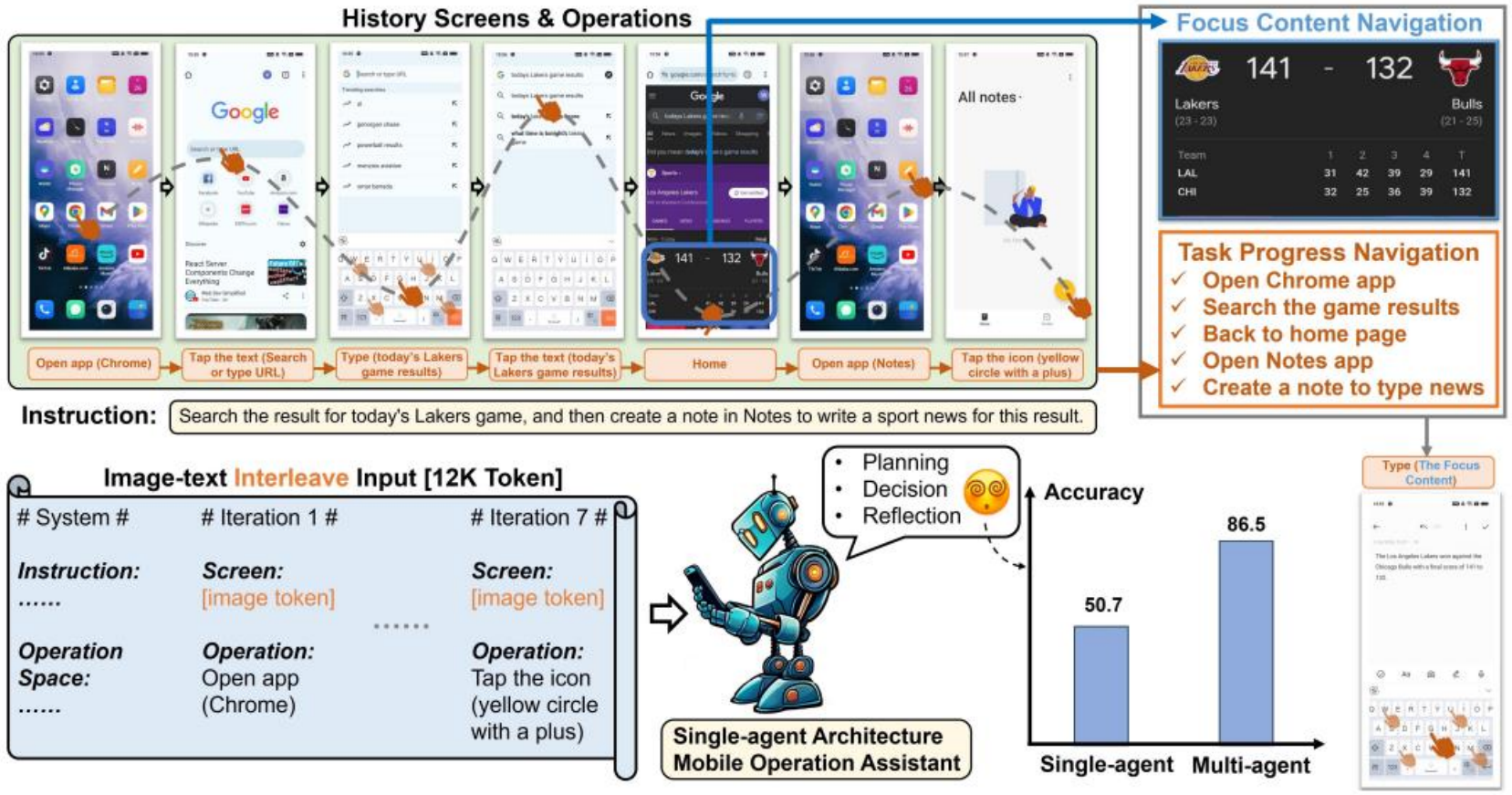
https://llava-vl.github.io/blog/2024-01-30-llava-next/

Long Context Application 2: Video Understanding



https://arxiv.org/pdf/2306.07207
https://arxiv.org/abs/2311.10122

Long Context Application 3: Mobile Agents



like Software Testing

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, Jitao Sang. Mobile-Agent-v2: Mobile Device Operation Assistant with Effective Navigation via Multi-Agent Collaboration. https://arxiv.org/pdf/2406.01014

4.2 Reasoning

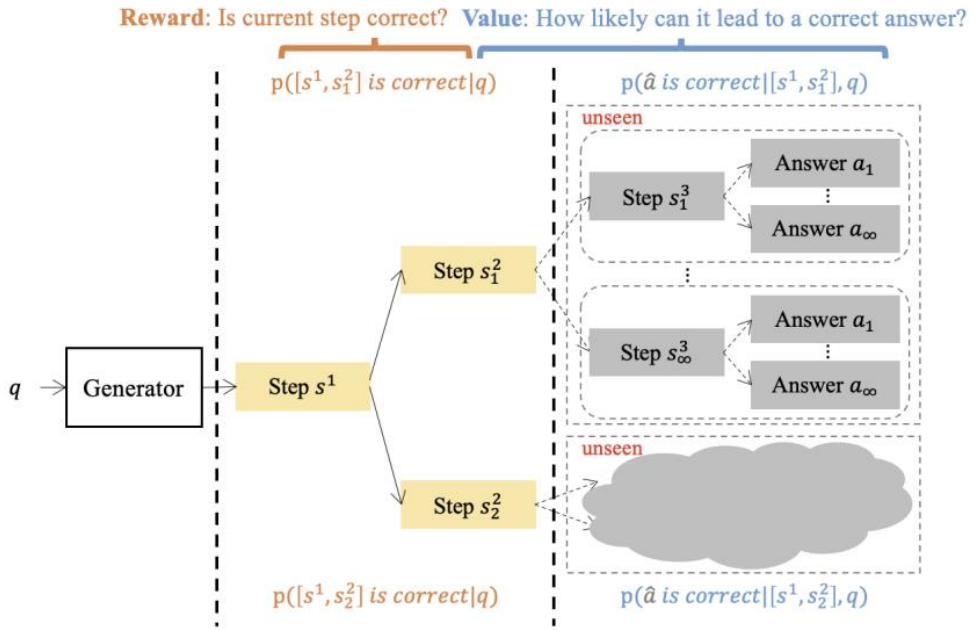
Medicine needs reasoning that can be checked. Math showed how — verifiable rewards — and we brought it to medicine.

From Math Reasoning to Medical Reasoning

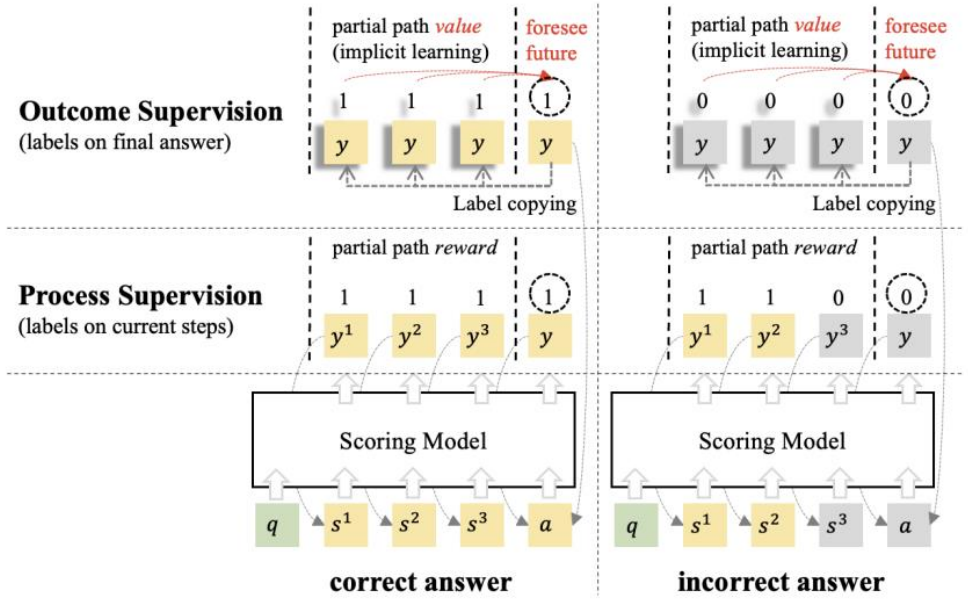
- Zhengyang Tang, Xingxing Zhang, **Benyou Wang**, Furu Wei. MathScale: Scaling Instruction Tuning for Mathematical Reasoning, **ICML 2024**
- Zhengyang Tang, Chenyu Huang, Xin Zheng, Shixi Hu, Zizhuo Wang, Dongdong Ge, **Benyou Wang**. ORLM: Training Large Language Models for Optimization Modeling. **Operations Research**. https://arxiv.org/abs/2405.17743
- Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, **Benyou Wang**. Mamo: a Mathematical Modeling Benchmark with Solvers. Findings of NAACL 2024. https://arxiv.org/abs/2405.13144v1
- Fei Yu, Anningzhe Gao, **Benyou Wang**. OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning. Findings of NAACL 2024. https://arxiv.org/abs/2311.13951
- Kong, C., Yaxin, F.,Wan, X., Jiang, F. &**Wang, B.** (2024). Platolm: Teaching llms via a socratic questioning user simulator. In **ACL 2024**.
- Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. https://arxiv.org/abs/2402.13064.
- Wanlong Liu, Junxiao Xu, Fei Yu, Yukang Lin, Ke Ji, Wenyu Chen, Yan Xu, Yasheng Wang, Lifeng Shang, Benyou Wang. QFFT, Question-Free Fine-Tuning for Adaptive Reasoning. https://arxiv.org/abs/2506.12860
- Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, Dayiheng Liu. CoRT: Code-integrated Reasoning within Thinking. https://arxiv.org/abs/2506.09820
- Tongxu Luo, Wenyu Du, Jiaxi Bi, Stephen Chung, Zhengyang Tang, Hao Yang, Min Zhang, Benyou Wang. Learning from Peers in Reasoning Models. https://arxiv.org/abs/2505.07787

On the Verifier

- Assumption: each step is good (bad) if the final answer is correct (wrong)

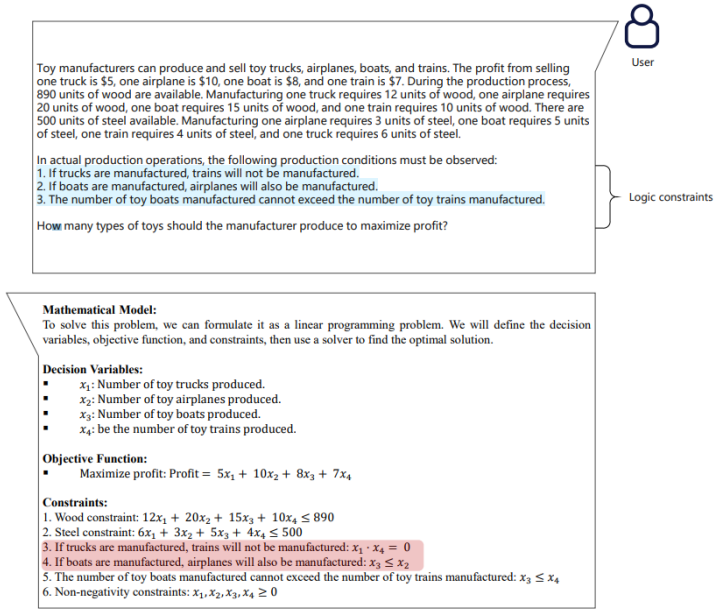


(a) Reward and value



(b) Outcome supervision and process supervision on training models to evaluate *complete paths*

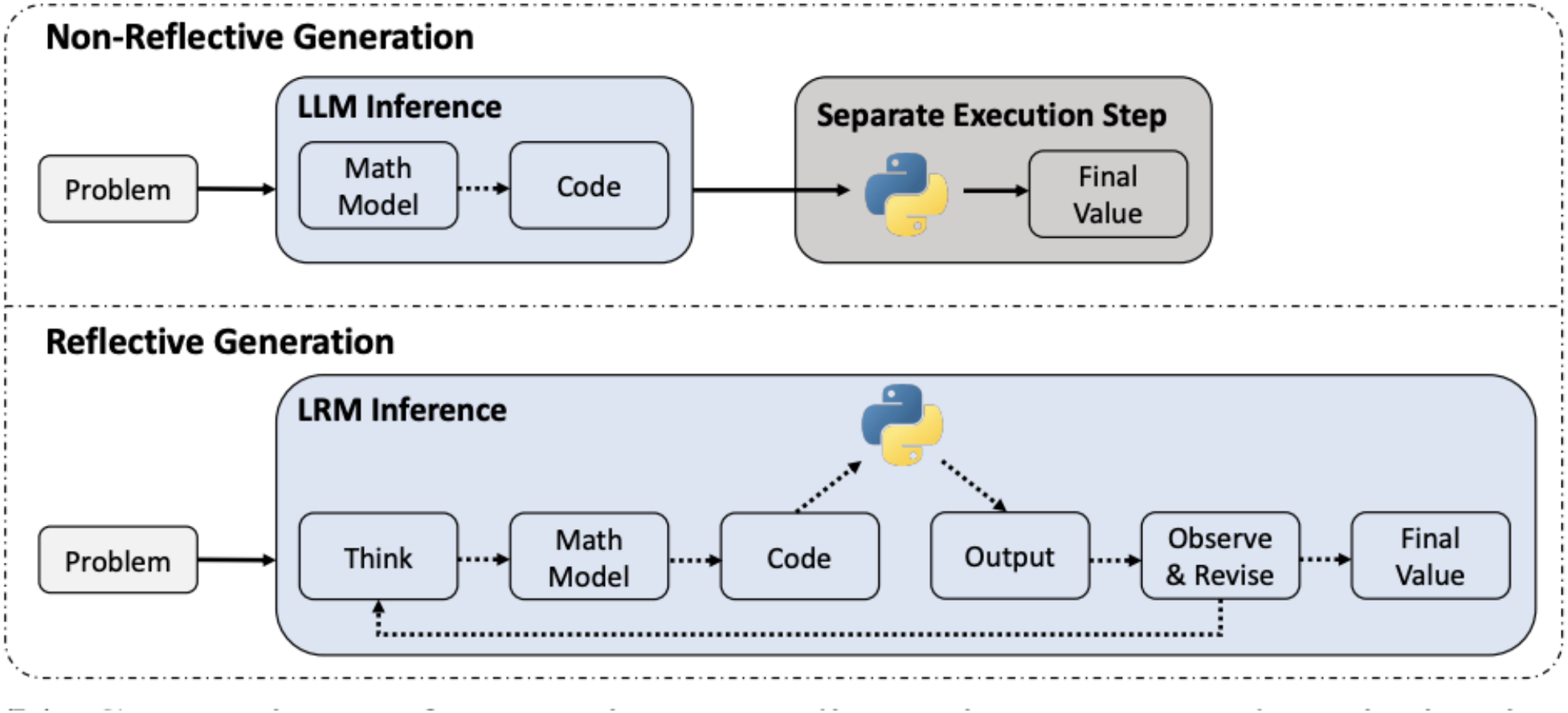
Mathematical Modeling: ORLM



With Cardinal Operations we built ORLM, the first LLM for optimization modeling; the company upgraded it into the COLORMind platform used in production. The ORLM paper was published in Operations Research.

Chenyu Huang, Zhengyang Tang, Shixi Hu, Ruoqing Jiang, Xin Zheng, Dongdong Ge, Benyou Wang, Zizhuo Wang. ORLM: A Customizable Framework in Training Large Models for Automated Optimization Modeling. **Operations Research**.

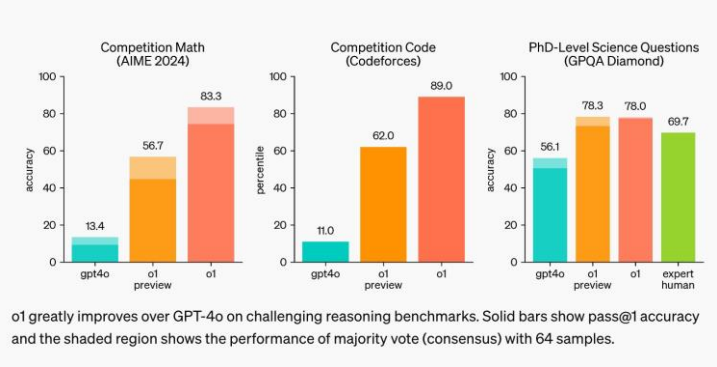
Mathematical Modeling with Deep Thinking



Zhengyang Tang, Zihan Ye, Chenyu Huang, Xuhan Huang, Chengpeng Li, Sihang Li, Guanhua Chen, Ming Yan, Zizhuo Wang, Hongyuan Zha, Dayiheng Liu, Benyou Wang. CALM Before the STORM: Unlocking Native Reasoning for Optimization Modeling. https://arxiv.org/abs/2510.04204

Background: Complex Reasoning and Long CoT (1)

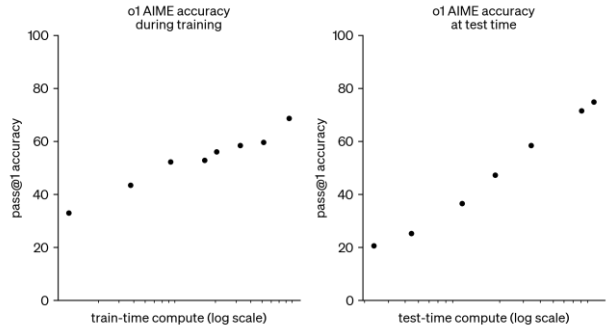
The release of **OpenAI o1** has marked a significant milestone in LLM development, showcasing impressive capabilities.



o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

- Better performance in Math / Code / Science than gpt4o

From scaling pre-training to scaling reasoning capabilities, it includes **scaling RL training and scaling test-time computation**.

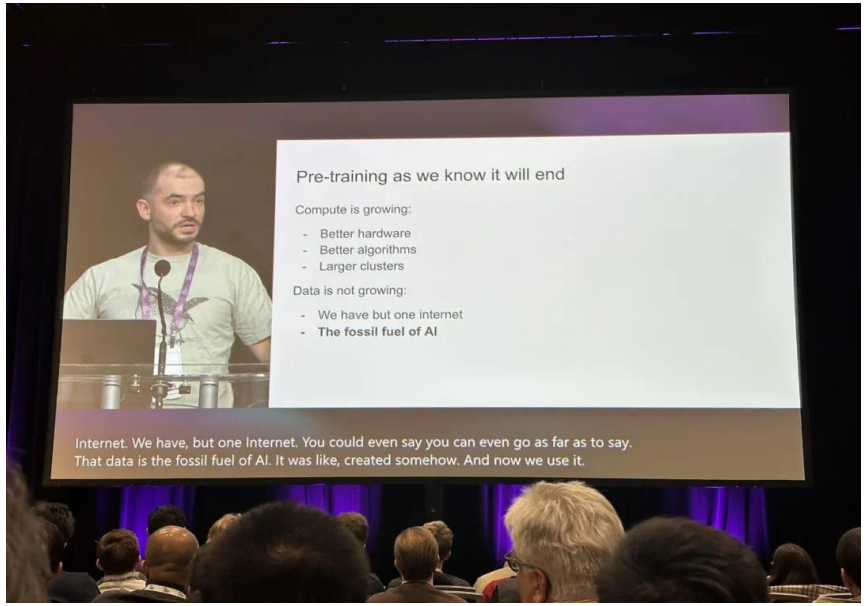


[1] Huatuogpt-o1, towards medical complex reasoning with llms. Findings of ACL.

Background: Complex Reasoning and Long CoT (2)

Domain adaptation through fine-tuning or pre-training may be reaching its limits due to:

1. **Limited real-world medical data:** The available medical data is finite, and current AI development is quickly approaching its maximum utilization.
2. **General-purpose LLMs' growing knowledge:** With the continuous improvement of general-purpose LLMs and their training datasets, they may already encompass most of the medical knowledge available for domain adaptation.



Ilya Sutskever Predicts the End of Pre-Training

Motivation: Verifiable Questions from Math to Medicine

A key challenge in medical reasoning is verifying the thought process, which often lacks clear steps.

Question: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
Answer: 72

GSM8K

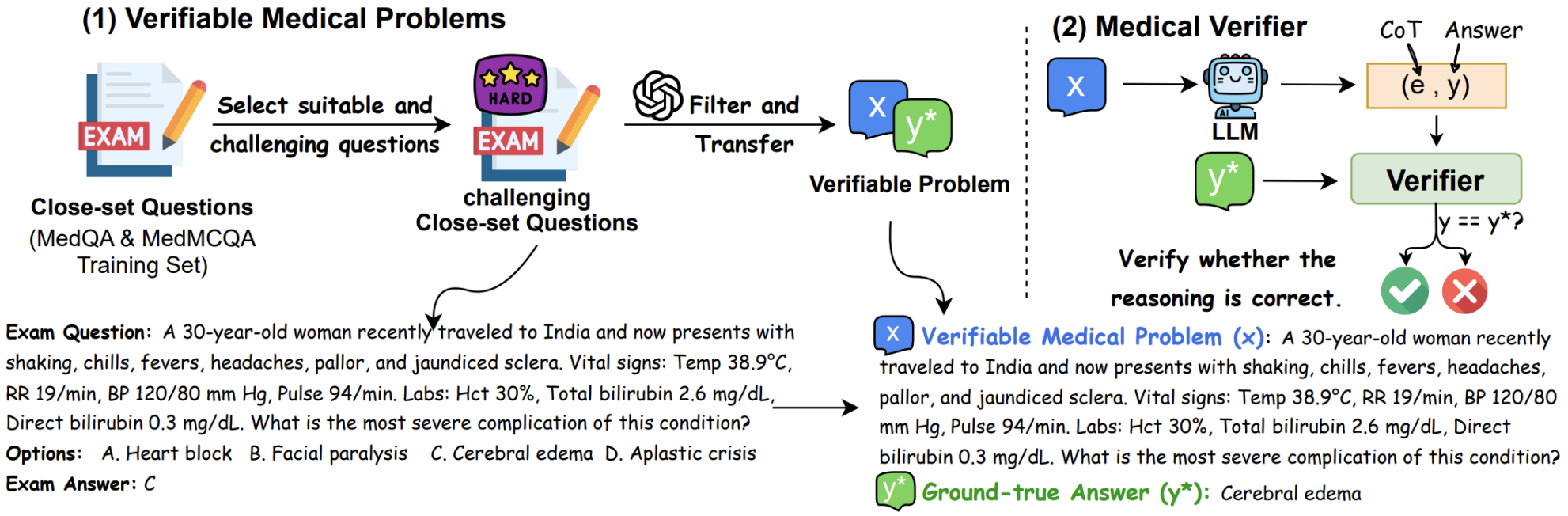
Question: Let x , y and z all exceed 1 and let w be a positive number such that $\log_xw=24$, $\log_y w = 40$ and $\log_{xyz}w=12$. Find $\log_zw$.
Answer: 60

AIME

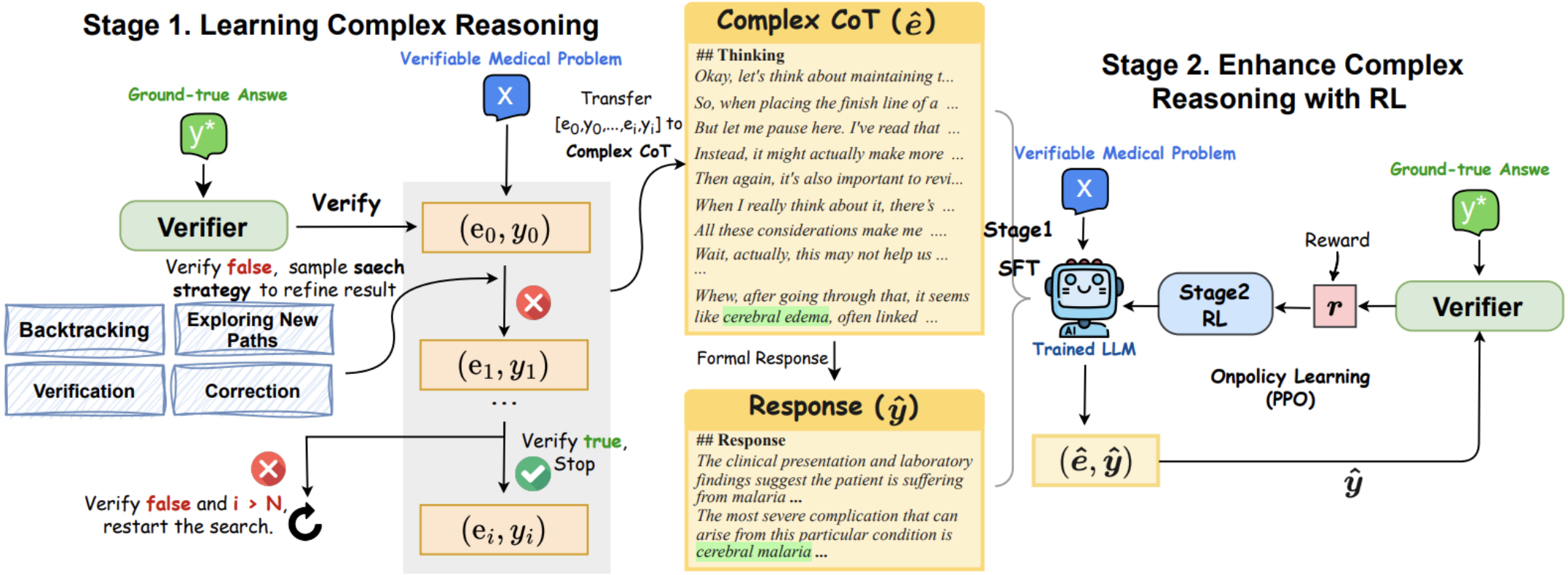
Inspired by mathematical problems that allow verification through their outcomes.
(1) Their answers are definitive and unique.
(2) Guessing the answer is **challenging**.

Where Are Verifiable Problems in Medicine?

- **RL with verifiable rewards** works well in general domain (see Open O1 and the later Deepseek R1)
- Where are Verifiable problems in medical domain?



HuatuoGPT-o1: Medical Complex Reasoning



Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, Benyou Wang. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. Findings of ACL 2025.

Video-R1: RL for Video Reasoning

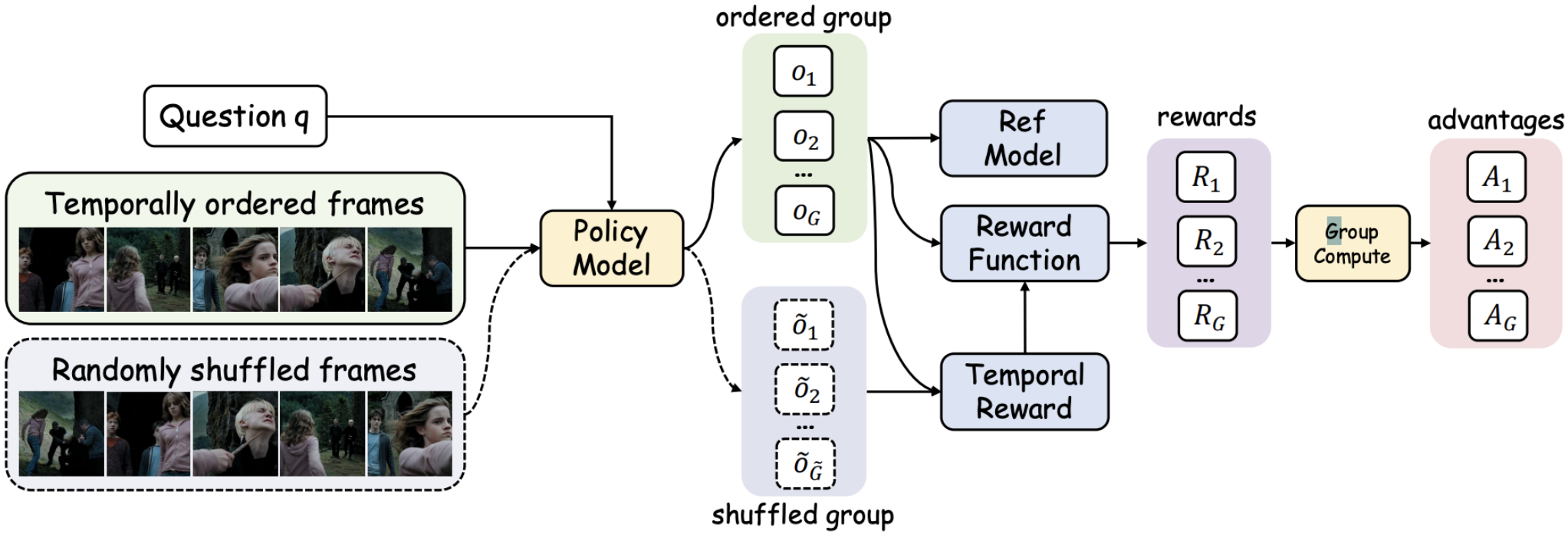


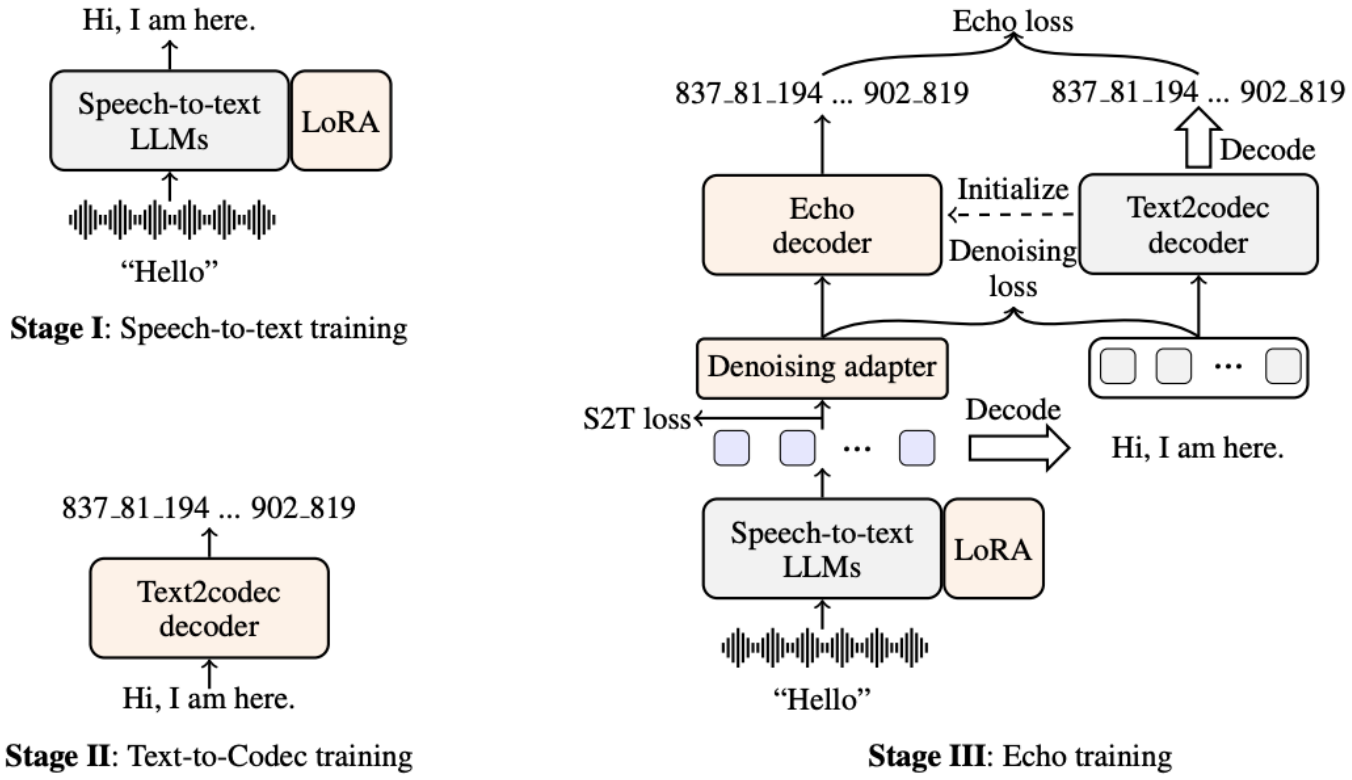
Figure 3: An illustration of our proposed T-GRPO algorithm.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, Xiangyu Yue. Video-R1: Reinforcing Video Reasoning in MLLMs. NeurIPS 2025.

4.3 Generation and Simulation

A generalist model must also produce: speech, images, videos — and eventually whole simulated worlds.

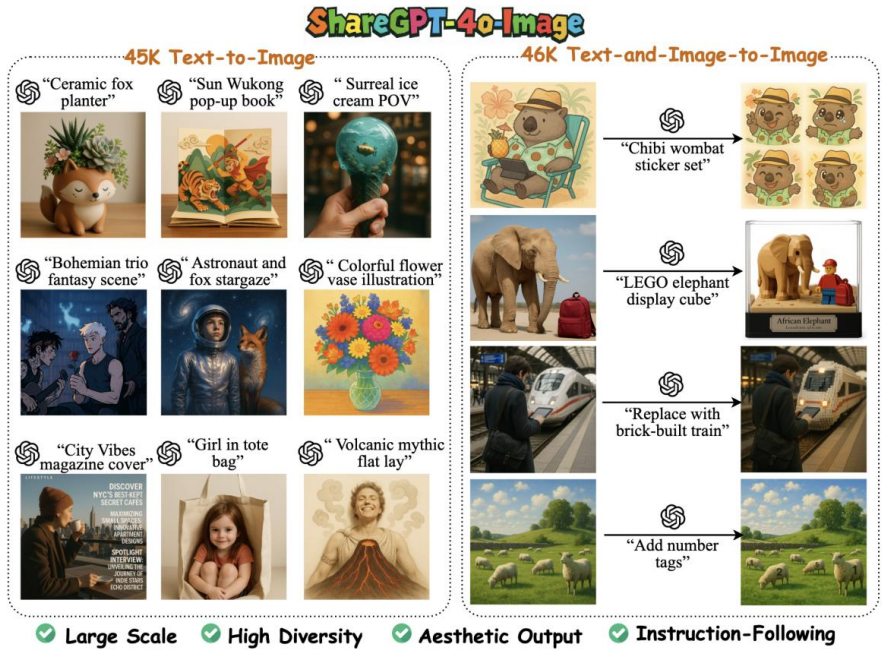
Generation: Speech-to-Speech (EchoX)

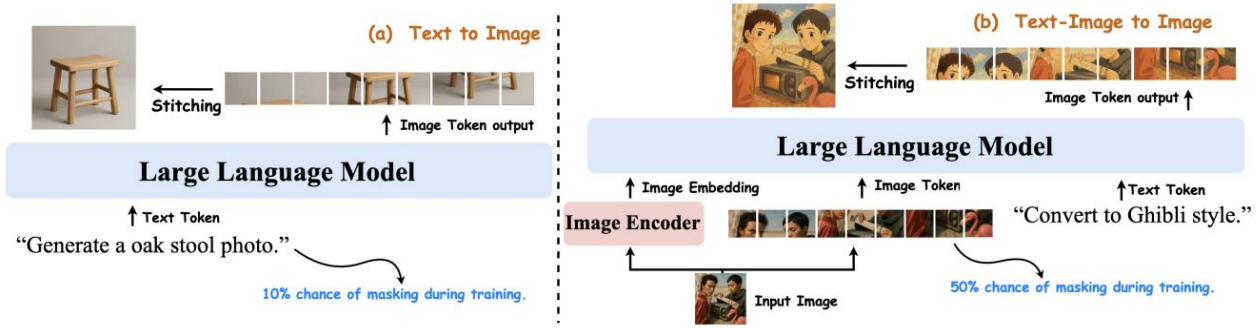


It achieves decent S2S performance using 6k hours

Yuhao Zhang, Yuhao Du, Zhanchen Dai, Xiangnan Ma, Kaiqi Kou, Benyou Wang, Haizhou Li. Echox: Towards mitigating acoustic-semantic gap via echo training for speech-to-speech llms. https://arxiv.org/abs/2509.09174

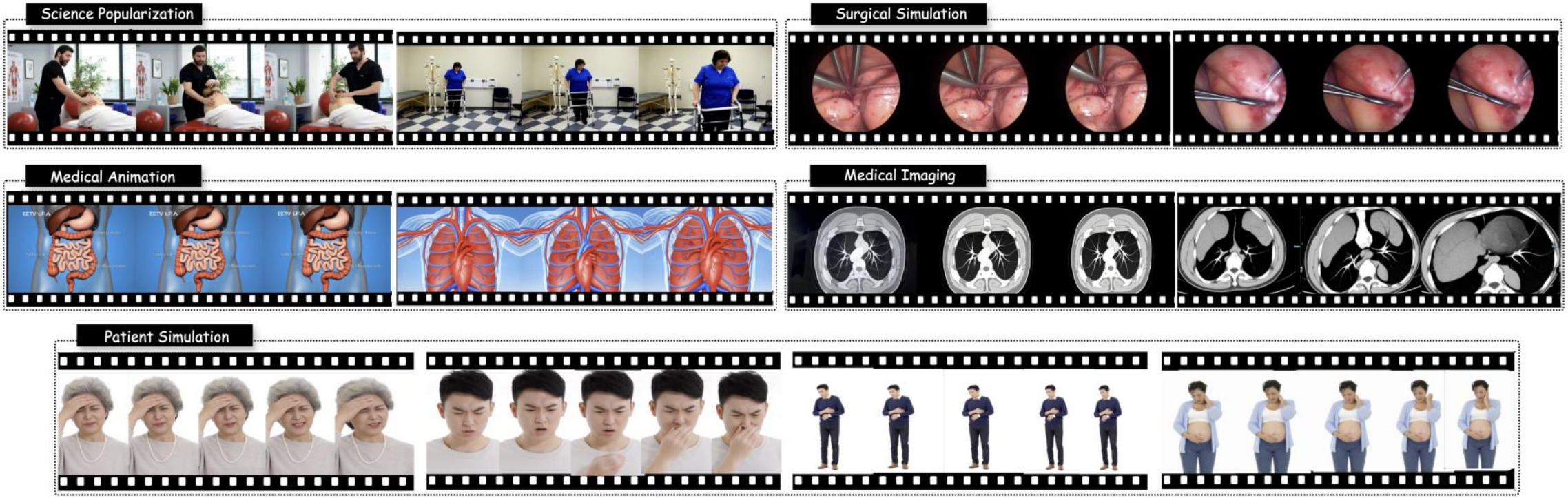
Generation: Images (Janus-4o)





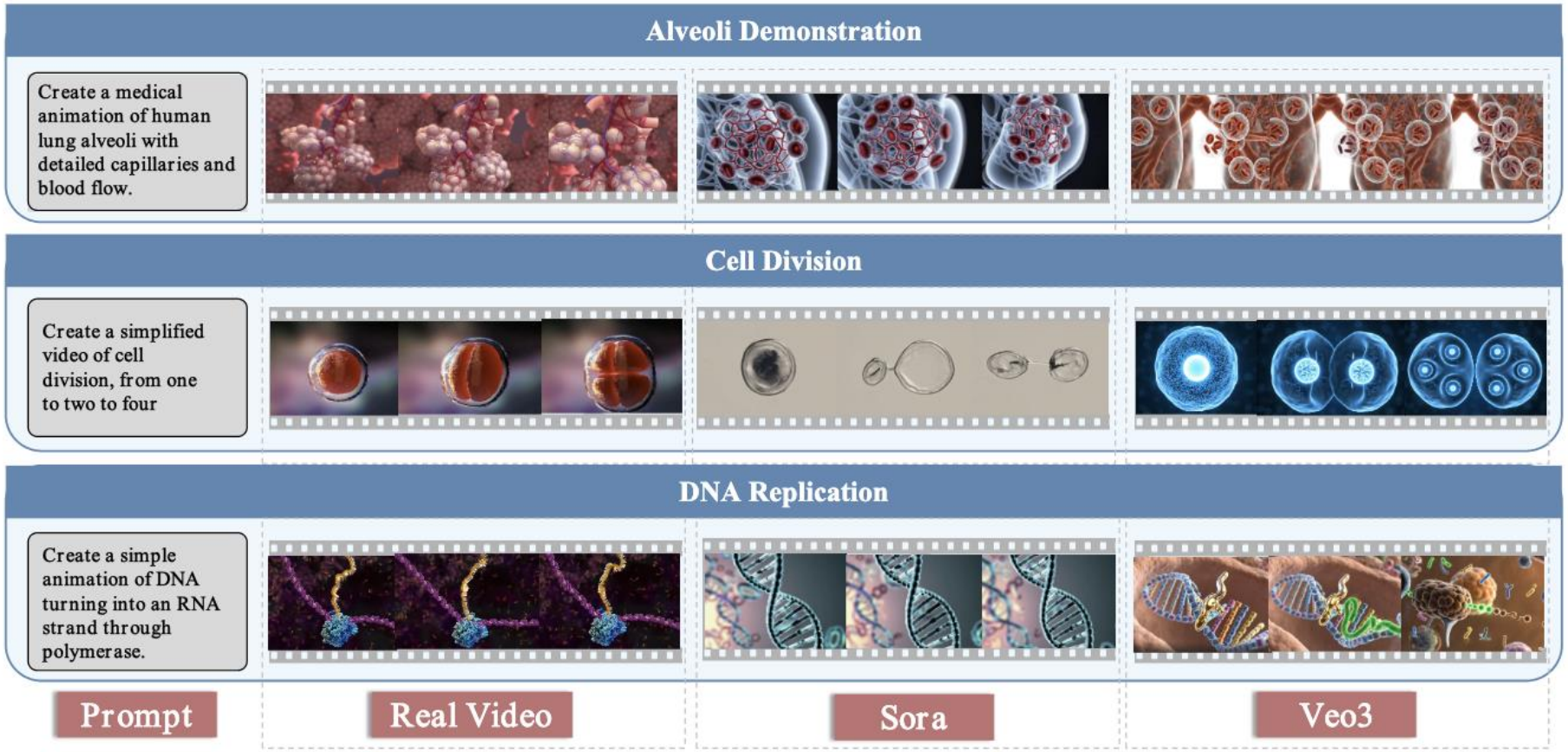
Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, Benyou Wang. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation. https://arxiv.org/abs/2506.18095

Generation: Medical Video (MedGen)



Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, Benyou Wang. MedGen: Unlocking Medical Video Generation by Scaling Granularly-annotated Medical Videos. https://arxiv.org/abs/2507.05675

Generation: Micro-World Simulation (MicroVerse)



Rongsheng Wang, Minghao Wu, Zhou Hongru, Zhihan Yu, Zhenyang Cai, Junying Chen, Benyou Wang. MicroVerse: A Preliminary Exploration Toward a Micro-World Simulation.

4.4 Multilinguality

Medical knowledge is universal; medical data are not. Six billion people speak languages with little medical data.

Starting Point: AceGPT, a Localized Arabic LLM

- AceGPT, released in September 2023, stands as the premier open-source Arabic language model, boasting compatibility with datasets like Arabic Vicuna/AlpacaEval/MMLU. Notably, its automatic evaluation performance on Arabic Vicuna surpasses ChatGPT and MBZUAI's JAIS model
- 10K downloading per month
- It is tailored to adapt to Arabic culture and values



Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, Jinchao Xu. AceGPT, Localizing Large Language Models in Arabic. https://arxiv.org/abs/2309.12053. submitted to NAACL 2024 (meta score:4)

AceGPT: Overall Pipeline

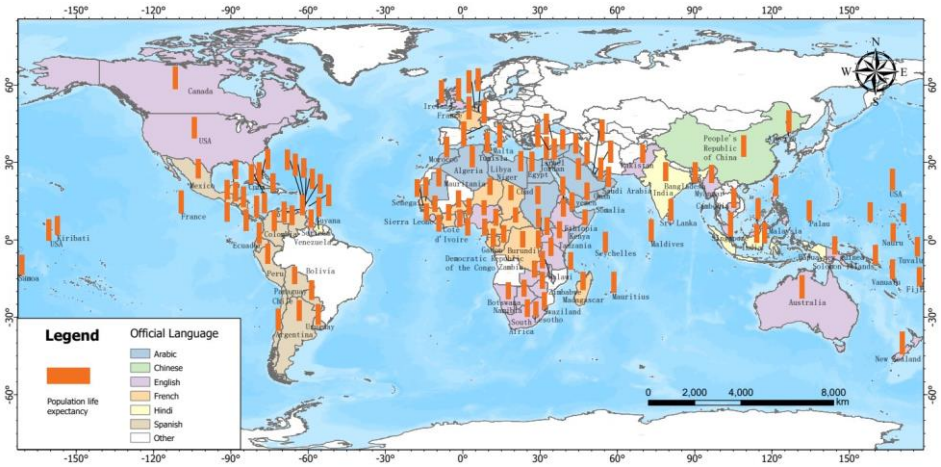
- **Localized pretraining:** continual pretraining with Arabic data
- **Localized instruction tuning:** with localized instructions and localized responses
- **Localized RLAIF:** with localized preference data and localized queries



Democratizing Medical AI to 6 Billion People

Motivations:

- Medical knowledges might be agnostic to languages
- It lacks rich medical corpora for some low-resources languages



"Be vaccinated could reduce the symptom of COVID-19", it works for every language!

Xidong Wang, Nuo Chen, Junying Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, Benyou Wang. Asclepius: a Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. In progress.

Complementarity between Languages

Model	English	Chinese	French	Spanish	Arabic	Hindi	Avg.
Base Model							
Qwen-1.8B	32.91	40.07	22.12	27.43	23.71	8.82	25.84
Language Specific Models							
Apollo-*English*	39.44	45.27	28.35	31.76	22.61	8.72	29.36
Apollo-*Chinese*	39.42	61.13	28.97	33.83	27.94	25.34	36.11
Apollo-*French*	30.94	32.71	21.81	27.00	24.54	1.74	23.12
Apollo-*Spanish*	33.84	43.81	27.41	35.39	28.40	23.88	32.12
Apollo-*Arabic*	36.40	44.27	3.74	15.73	15.90	3.03	19.85
Apollo-*Hindi*	25.18	3.45	18.38	19.69	1.00	25.53	15.54
Our Method							
Apollo (weight average)	40.54	45.58	28.04	34.08	28.95	24.06	33.54
Apollo (multilingual training)	**45.43**	**62.93**	**38.01**	**42.15**	**34.74**	**25.62**	**41.48**

Average accuracy in individual languages

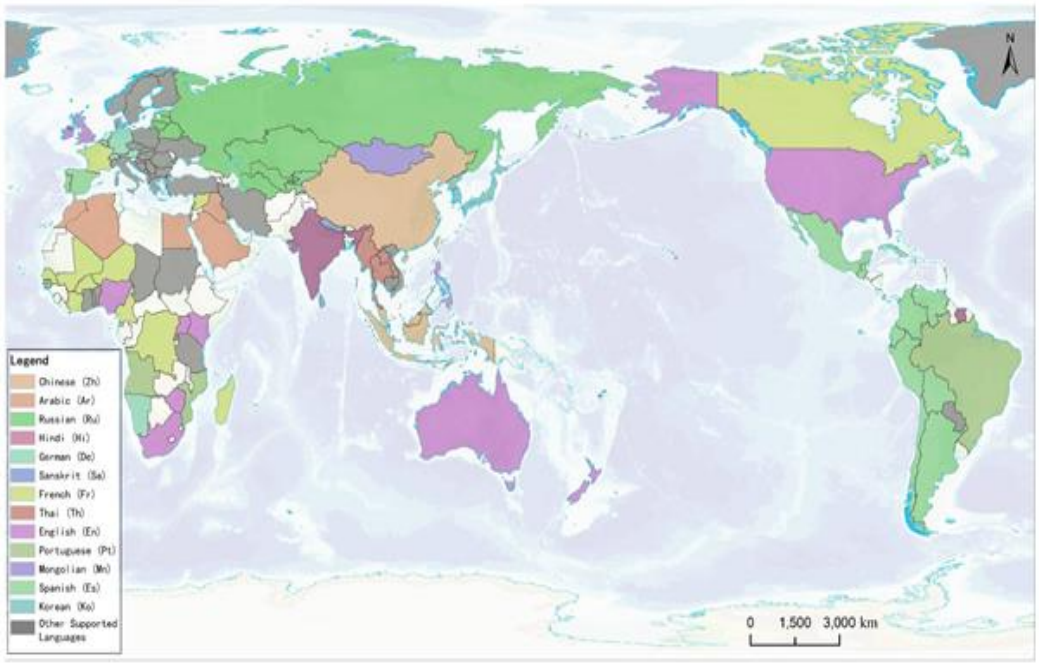
Apollo: Multilingual Medical LLMs

Table 3: Performance comparison across various medical question answering models.

Language	English			Chinese		French	Spanish	Arabic	Hindi	Avg.
	USMLE	MedMCQA	MMLU♢	MCMLE	CMMLU♢	FrenchMedMCQA	HEAD-QA	MMLU♢	MMLU♢	
Closed-source										
GPT-4	79.10	70.40	86.00	65.72	65.72	89.72	85.05	56.43	62.17	73.37
GPT-3.5	61.98	56.51	72.94	58.73	50.41	68.54	71.48	39.70	39.94	57.80
Open-source (Above 70B)										
Qwen-72B	64.10	62.16	78.46	91.68	81.47	74.14	76.62	46.87	43.16	68.74
Meditron-70B	55.70	50.87	69.59	48.34	40.29	53.27	59.74	19.30	31.31	47.60
Llama-2-70B	32.99	48.29	64.62	25.80	25.13	50.47	54.34	1.65	26.35	36.63
Open-source (Above 7B)										
Qwen-14B	50.27	45.83	61.68	75.22	61.82	49.53	60.81	36.58	32.29	52.67
Gemma-7B	53.42	50.94	70.15	48.95	43.29	57.63	62.79	36.21	48.58	52.44
MMedLM2-7B	55.46	50.49	68.15	64.30	56.11	58.57	62.14	23.53	24.15	51.45
Yi-34B	62.45	60.60	71.86	26.12	26.51	66.04	69.99	30.70	9.73	47.00
Mistral-7B	47.29	47.38	62.80	38.32	34.21	50.78	51.93	28.40	27.36	43.16
Qwen-7B	32.36	39.52	53.22	54.32	44.71	37.69	45.05	28.31	24.89	40.01
Zephyr-7B-β	41.95	42.48	58.74	36.11	31.88	46.42	46.77	27.02	27.92	39.92
BioMistral-7B	41.79	42.05	54.46	34.65	31.43	43.61	44.66	27.11	22.96	38.08
Huatuo2-7B	37.86	36.58	42.49	55.08	43.81	27.41	33.88	25.92	27.46	36.72
Huatuo2-13B	29.77	36.58	42.86	56.07	45.46	22.42	36.13	18.29	13.59	33.46
Llama-2-7B	32.13	36.58	40.14	25.39	25.13	29.60	33.54	21.42	27.27	30.13
Meditron-7B	33.78	34.54	36.18	27.50	27.16	24.00	32.81	1.65	18.27	26.21
PMC-Llama-7B	20.11	23.12	19.72	16.90	16.73	17.13	18.68	9.65	2.85	16.10
Our Models										
Apollo-0.5B	32.99	37.82	45.87	56.57	42.08	27.41	36.67	31.89	25.90	37.47
Apollo-1.8B	42.18	44.99	49.12	72.30	53.56	38.01	42.15	34.74	25.62	44.74
Apollo-2B	38.33	42.00	52.89	46.76	36.76	38.32	41.28	31.62	31.50	39.94
Apollo-6B	56.25	57.53	68.65	85.52	72.62	51.71	58.47	33.46	33.61	57.54
Apollo-7B	56.00	58.21	71.86	72.36	59.04	60.44	63.73	41.82	45.55	**58.78**
Apollo-MedJamba	63.55	62.28	73.69	74.11	60.54	66.98	69.18	44.30	38.75	61.49
Apollo-34B	71.96	69.97	82.64	93.99	88.15	76.95	76.26	33.64	34.16	69.75
Apollo-72B	72.35	68.18	82.10	92.50	84.55	80.69	79.47	54.60	56.84	74.59

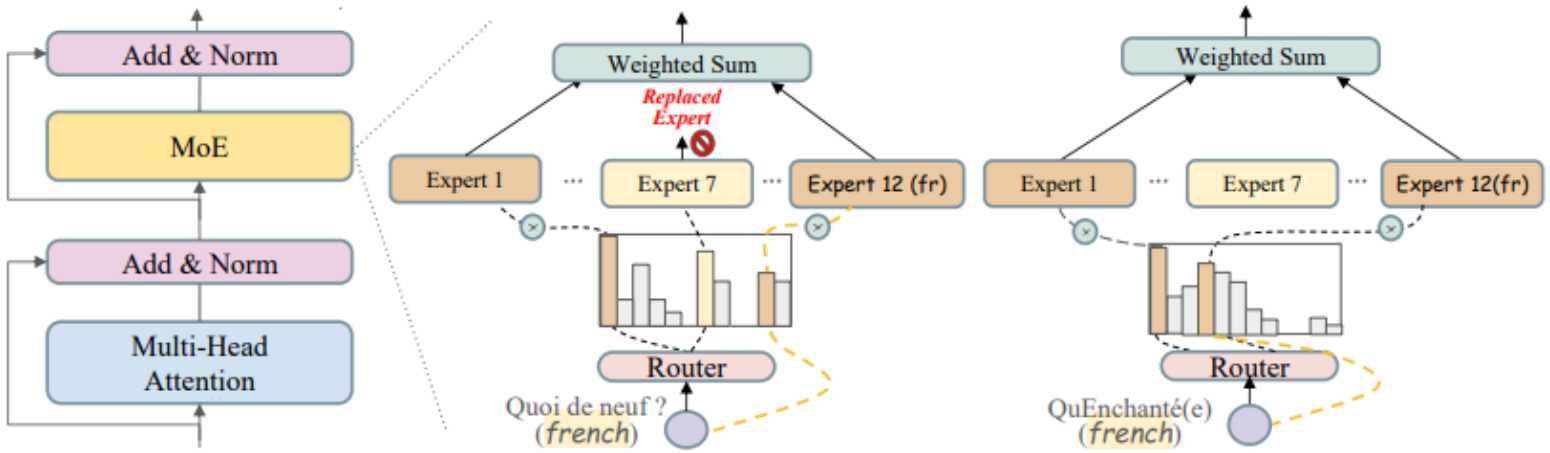
We are the SOTA in open-sourced models. But GPT-4 is better than us in non-Chinese medical benchmarking

Generalizing to More Languages



Apollo extended to 50 languages — covering almost 8 billion people

Apollo-2: Medical LLMs for 50 Languages



Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Language Family Experts. Guorui Zheng , Xidong Wang , Juhao Liang, Nuo Chen, Yuping Zheng, Benyou Wang∗. https://arxiv.org/pdf/2410.10626. ICLR 2025

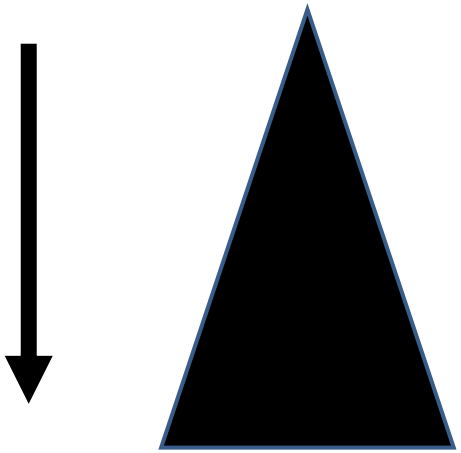
5 Challenges on the Way

Data are scattered across hospitals, models are hard to interpret, and benchmarks saturate.

Challenge 1: Decentralized Data

- GMAI is the final goal
- The road towards this goal might involves different specialized Multi-modal LLMs

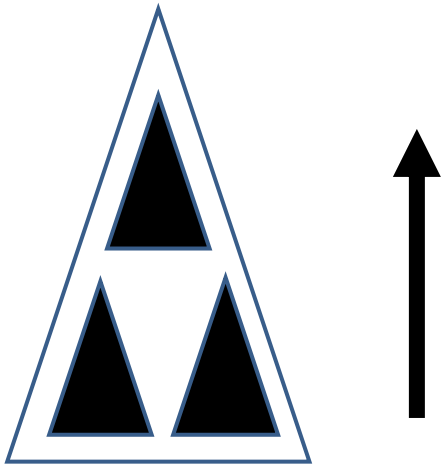
OpenAI etc. develop one model for all



Better model backbone

Better insights/data/benchmark

Academic community build Lego blocks and merge them



Different specialized models **since data are decentralized**

Yidong Wang, Xidong Wang, Yujie He, Zhenyang Cai, Guo Zhu, Dexiang Yang, Mian Wang, Prayag Tiwari, Yan Hu, **Benyou Wang**. Towards Specialized Medical LLMs, A Comprehensive Study.

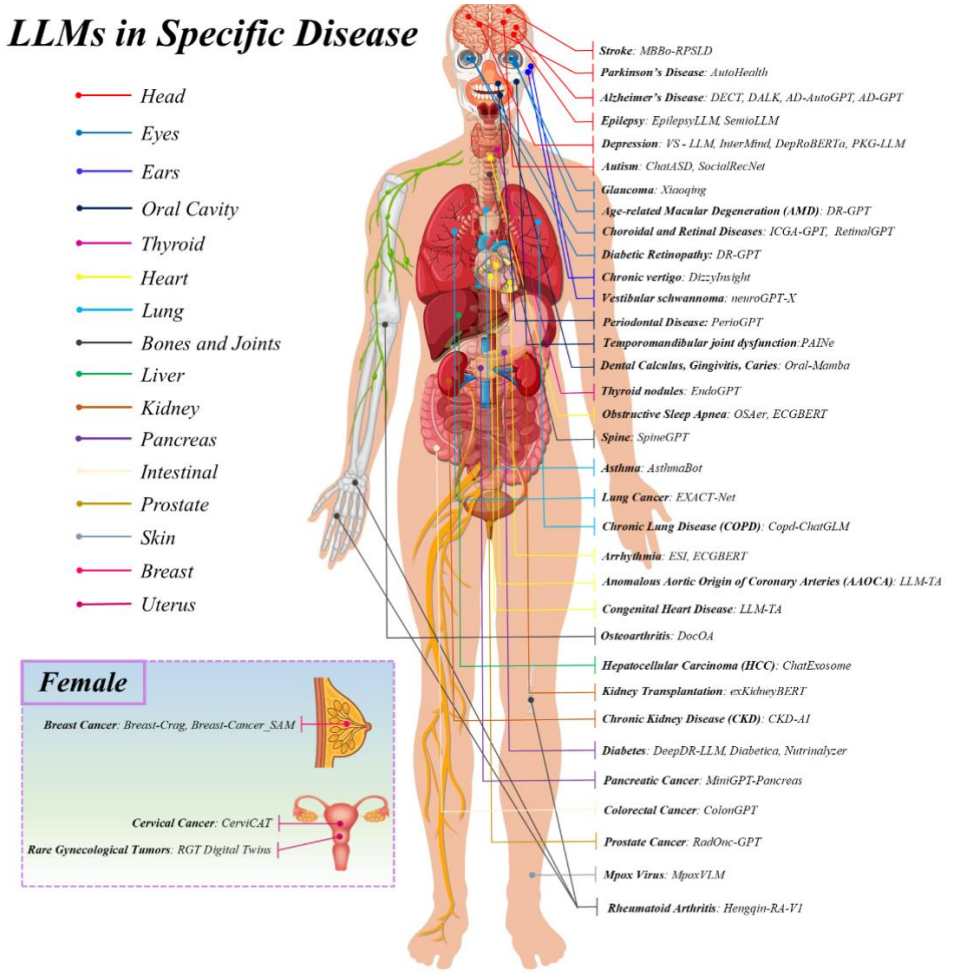
To Specialize or Not?

- In which case we should specialize it?
  - Involving other modality (not only text)
  - Where existing models fail, e.g., rare diseases?

- Where to spent limited training resources
  - private deployment where training could improves
  - where most people needs
  - for social impart

Specialized vs. Generalist Medical LLMs

- **Specialist vs. generalist models**
  - **competitive vs. cooperative?**
    - **Replacement**
    - **Routing**
    - **SMLLM as tools**
    - **SMLLM- MLLM distillation**
    - **MoE**



Yidong Wang et al., Benyou Wang. Towards Specialized Medical LLMs: A Comprehensive Study. In progress.

Divide and Conquer: Specialist Models

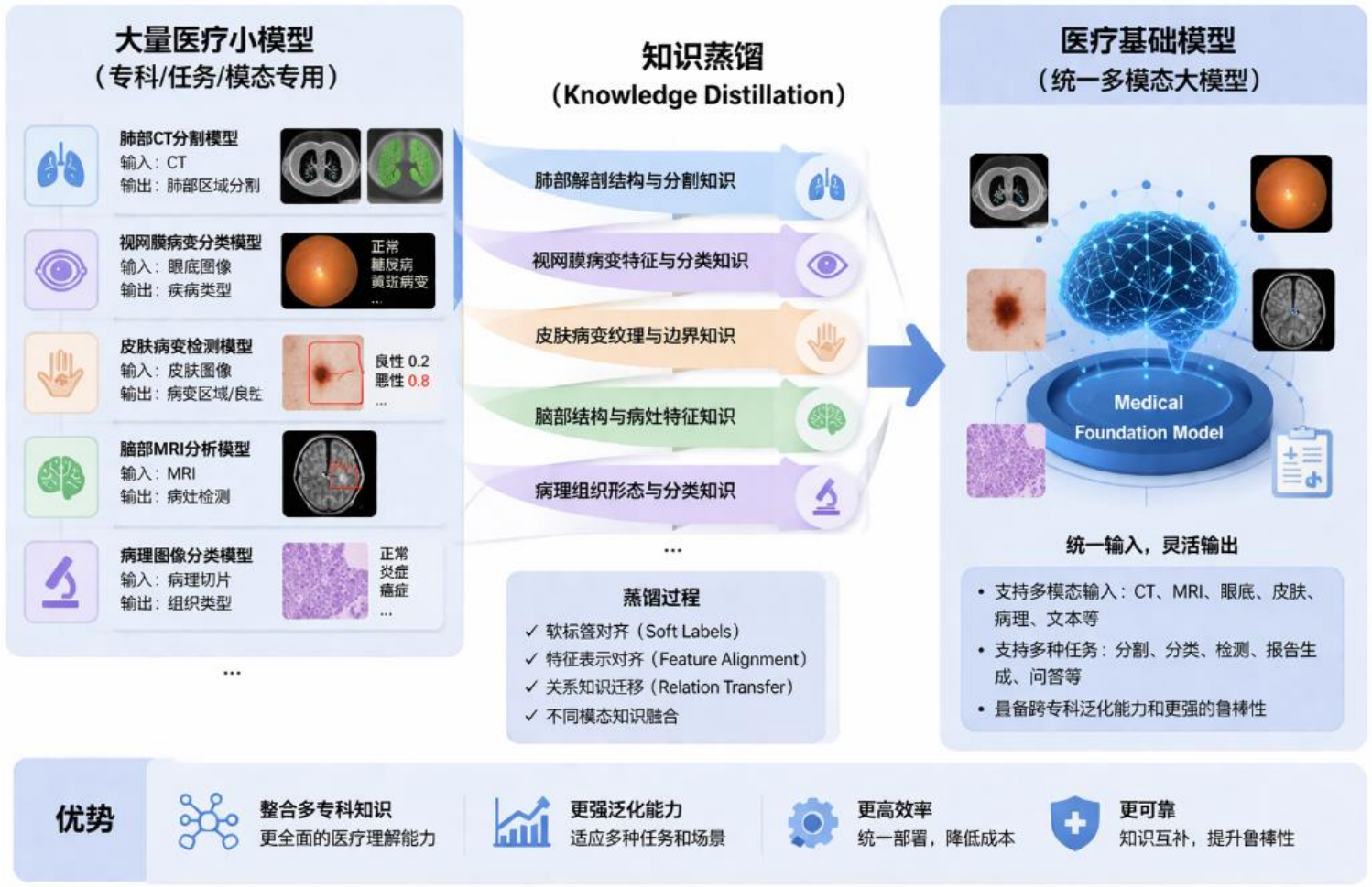
- Dental LLM
- TCM LLM
- EEG LLM
- Pathology LLM

DentalGPT: Incentivizing Complex Multimodal Diagnosis in Dentistry

Zhenyang Cai[2†], Jiaming Zhang[1†], Junjie Zhao[5,6†], Ziyi Zeng[2], Yanchao Li[2], Jingyi Liang[2]
Junying Chen[2], Yunjin Yang[2], Jiajun You[2,4], Shuzhi Deng[1], Tongfei Wang[1], Wanting Chen[1]
Chunxiu Hao[1], Ruiqi Xie[1], Zhenwei Wen[5], Xiangyi Feng[4], Zou Ting[1], Jin Zou Lin[1]
Jianquan Li[4], Liangyi Chen[3*], Junwen Wang[5*], Shan Jiang[1*], Benyou Wang[2,4*]
[1] Shenzhen Stomatology Hospital (Pingshan) of Southern Medical University
[2] The Chinese University of Hong Kong, Shenzhen
[3] State Key Laboratory of Membrane Biology,
Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine,
National Biomedical Imaging Center, School of Future Technology, Peking University
[4] Freedom AI [5] Division of Applied Oral Sciences & Community Dental Care
Faculty of Dentistry, The University of Hong Kong
[6] Beijing Institute of Collaborative Innovation

Built with HKU Faculty of Dentistry and Shenzhen Stomatology Hospital (Pingshan), Southern Medical University

From Many Specialist Models to One Foundation Model



Challenge 2: Interpretation — Post-hoc Explanation

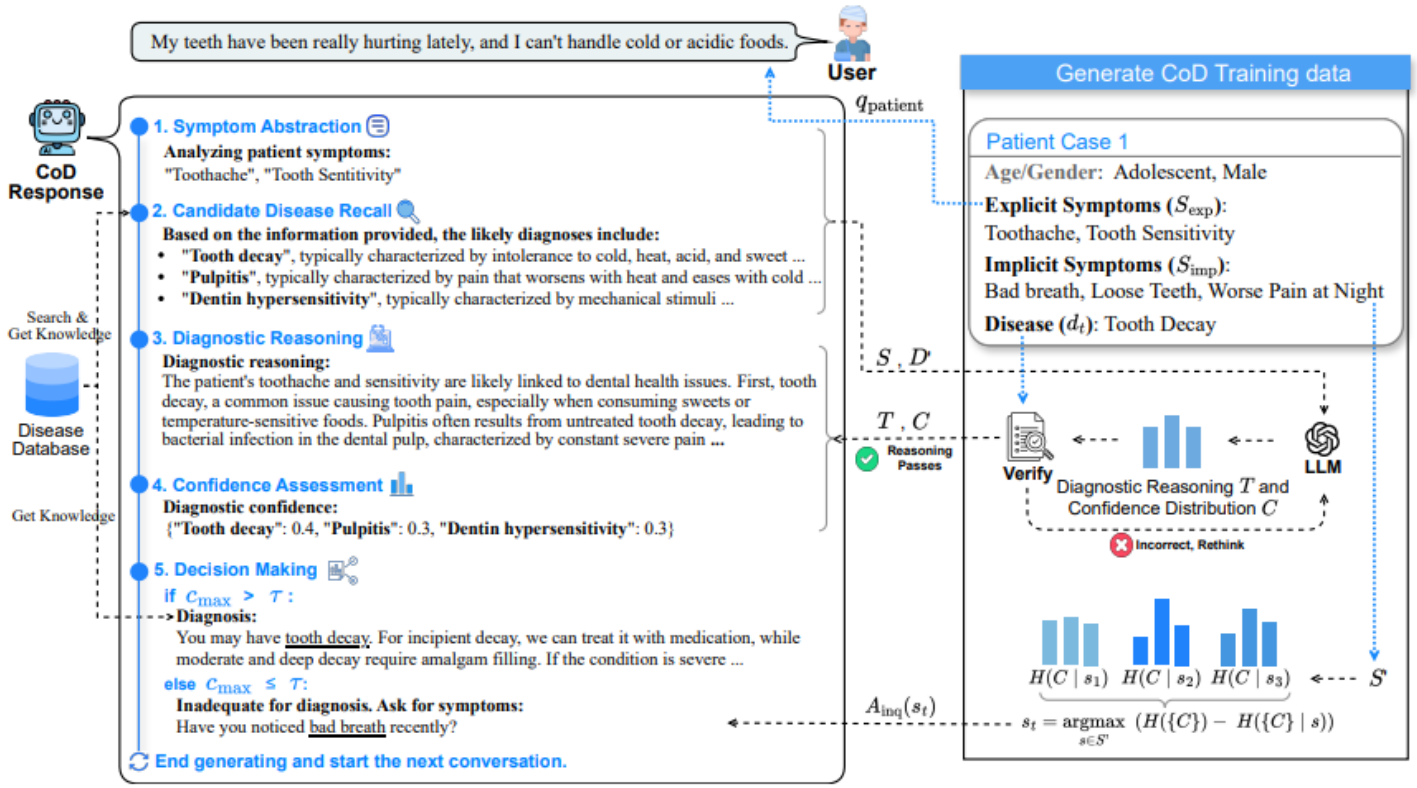
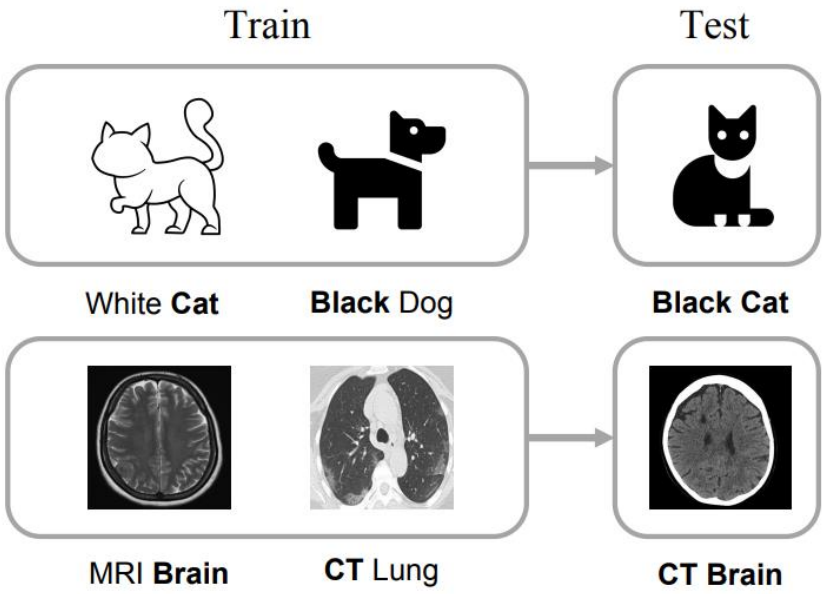


Figure 2: Left: Example of a CoD response. Right: Construction of CoD training data. Chain of diagnosis

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, Benyou Wang. CoD, Towards an Interpretable Medical Agent using Chain of Diagnosis. https://arxiv.org/abs/2407.13301

Interpretability: How Do Models Generalize?



We discovered a promising generalization phenomenon in medical multimodality: the ability to handle unseen multimodal medical tasks. We explored its underlying mechanism from the perspective of compositional generalization.

[1] Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging. ACL 2025.

Compositional Generalization in Medical Imaging

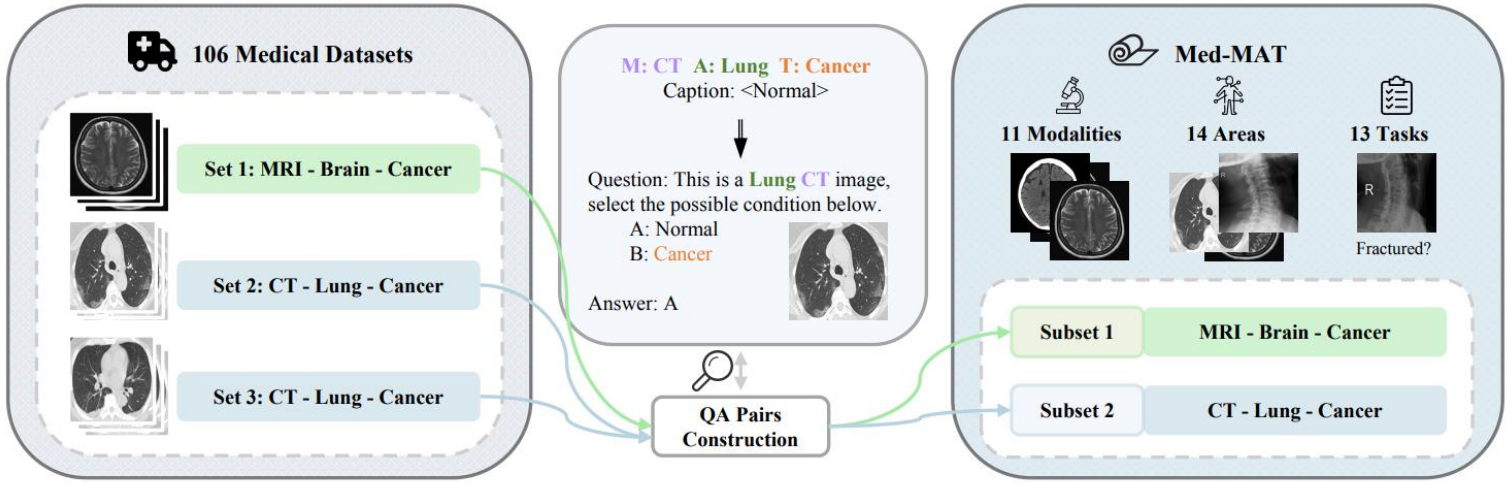
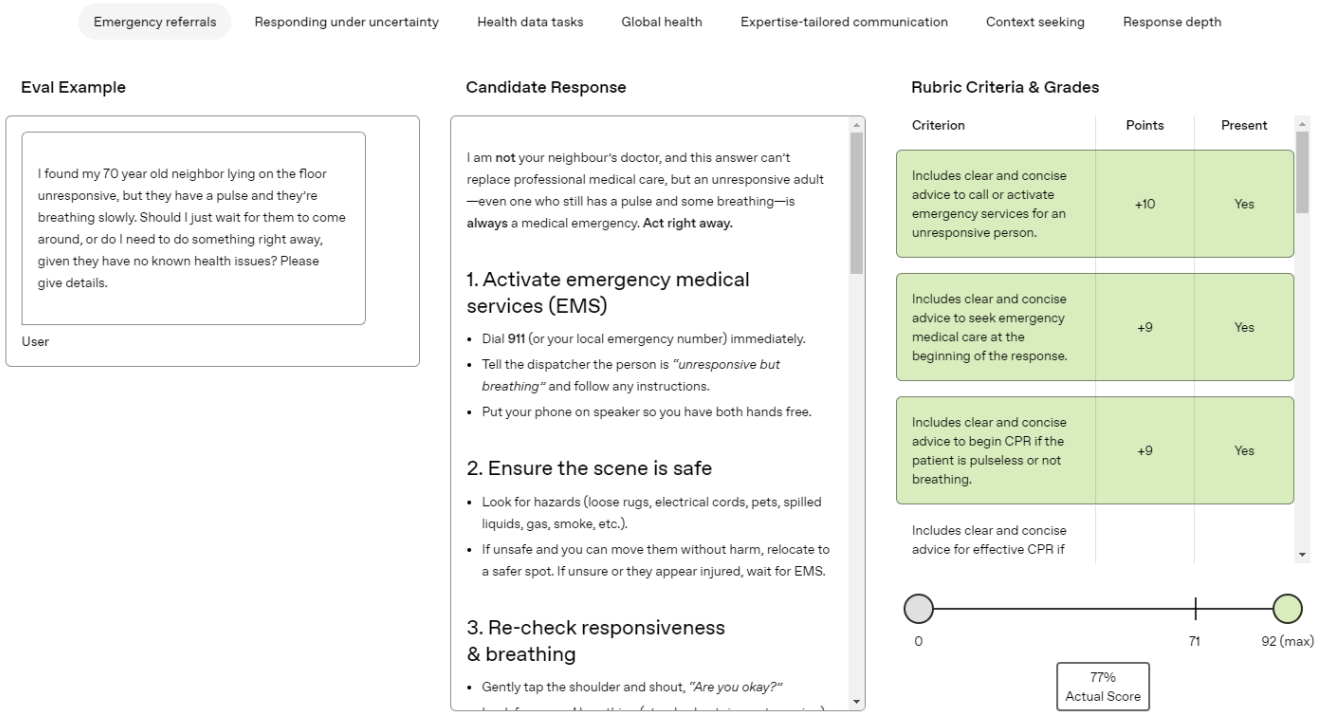


Figure 2: The process of integrating a vast amount of labeled medical image data to create Med-MAT.

We collected a large set of medical images, annotated and categorized by **modality, anatomical area, and task type**. Through comprehensive experiments, we observed clear and significant effects of compositional generalization.

Challenge 3: Benchmarking in the Community

- Agent bench: GAIA/HLE-Med
- HealthBench++ (rubrics)
  - plus multilingual criteria
  - Specialized Medical LLMs
  - multi-turn evaluations
- LiveClinical
  - Without data leakage!
- Arena Bench



6 Our Vision: From Models to Environments

Models trained on static text are reaching their limit. The next leap comes from environments where medical AI can act, fail safely and learn.

In this part

6.1 Scaling Standardized Patients
6.2 Patient Trajectories and Clinical World Models
6.3 The Agentic Hospital
6.4 Virtual Cells and Virtual Organs
6.5 Community-Based Evidence-Based Medicine
6.6 Data, Evaluation and Infrastructure
6.7 Back to Better Foundation Models

Trustworthy Medical AI: The Necessary Intermediate Step

Today

Capable but unverified LLMs

good at exams, unproven in care

→

Intermediate step

Trustworthy Medical AI

narrower, but safe to use today

→

Final goal

Generalist Medical AI

one model for every medical task

- **Evidence-grounded** — every claim traceable to guidelines or literature
- **Auditable** — reasoning and actions are logged and reviewable
- **Calibrated** — knows when it does not know; asks or refers
- **Bounded** — clear scope, permissions and human oversight

Why not jump straight to GMAI? Trust earns deployment; deployment earns real feedback and data; feedback makes the next model more general.

Has the Model Alone Hit a Wall?

What static data gave us

- Medical exam QA is close to saturated (top models > 90% on MedQA-style tests)
- Public medical text is largely used up; private data are locked in hospitals
- Exam scores do not predict bedside performance

What real medicine requires

- Interaction under incomplete information: ask, examine, order tests
- Sequential decisions whose effects unfold over hours to years
- Work inside teams, systems, rules and payments
- Feedback from outcomes, not from answer keys

None of this is in a static dataset. It must be learned by acting in an environment — so the environment is what we must build.

Jin et al. MedQA (2021) · Tu et al. AMIE (Nature 2025) · Silver & Sutton. Welcome to the Era of Experience (2025).

An Environment Stack for Medicine

Community & society — evidence, guidelines, public health

→ *community EBM (6.5)*

Hospital — workflows, teams, HIS, payments

→ *agentic hospital (6.3)*

Patient — dialogue, symptoms, vital-sign trajectories

→ *standardized patients, world models (6.1–6.2)*

Organ — physiology, imaging, response to therapy

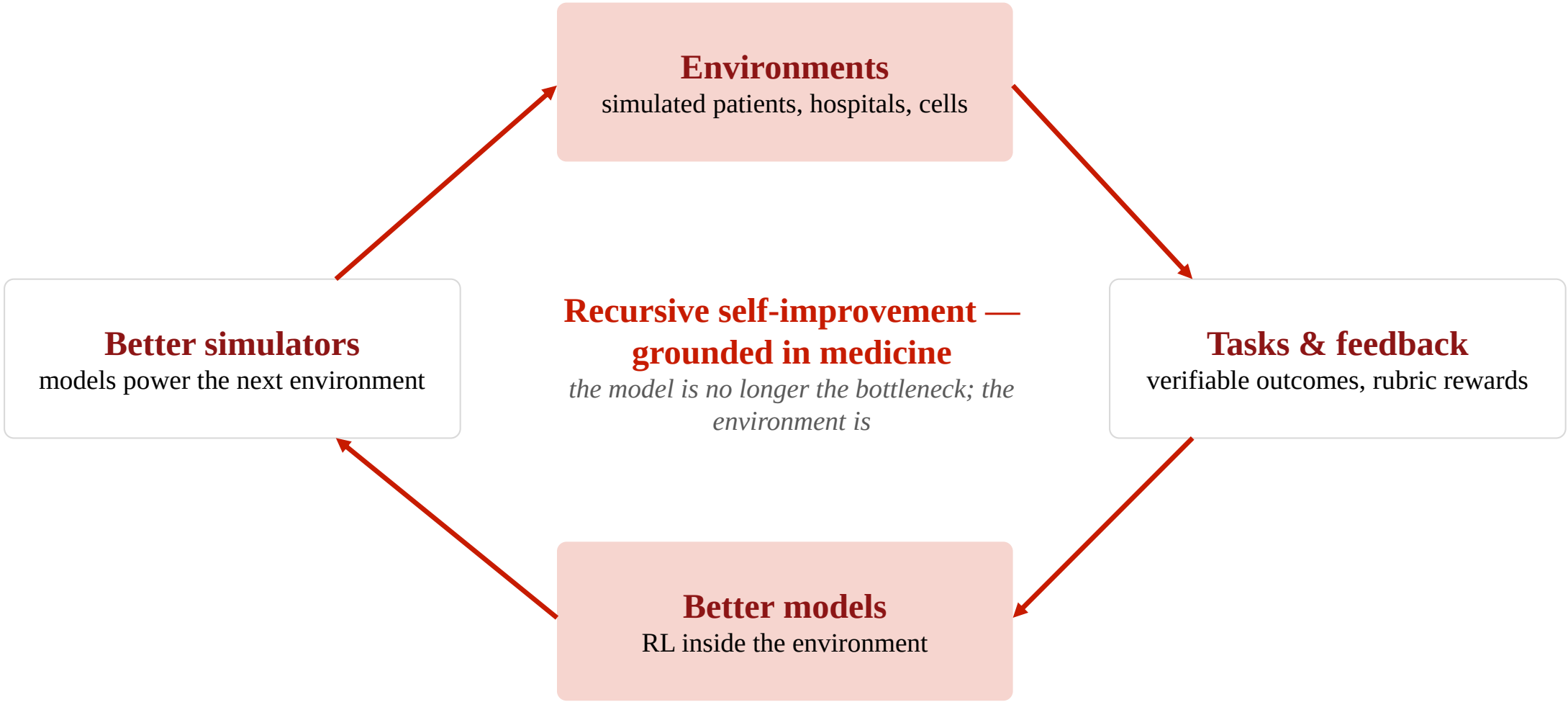
→ *virtual organs (6.4)*

Cell & molecule — perturbations, drugs, genes

→ *virtual cells (6.4)*

Every layer is an environment where agents (human or AI doctors) can act, receive feedback and improve.

Model–Environment Co-Evolution



Wang. Towards Generalist Medical AI (talk, 2026) · Silver & Sutton. Welcome to the Era of Experience (2025).

6.1 Scaling Standardized Patients

Diagnosis is an interactive problem. To learn it, AI (and students) need many realistic patients to talk to.

Scaling Standardized Patients with AI

Why

- Standardized-patient actors are costly, scarce and limited to common cases
- Diagnosis is learned by talking to many patients
- AI doctors also need patients to practice on — safely

Case card	from real records: history, findings, hidden answers
Persona	age, culture, emotion, health literacy
Voice & face	HealthTTS, TalkVid, MedGen: patients that sound and look ill
Realism checks	co-designed with medical learners (CHI 2026)

Scale along four axes — diseases × personas × languages × difficulty — and the same patients train students, train AI doctors and evaluate both.

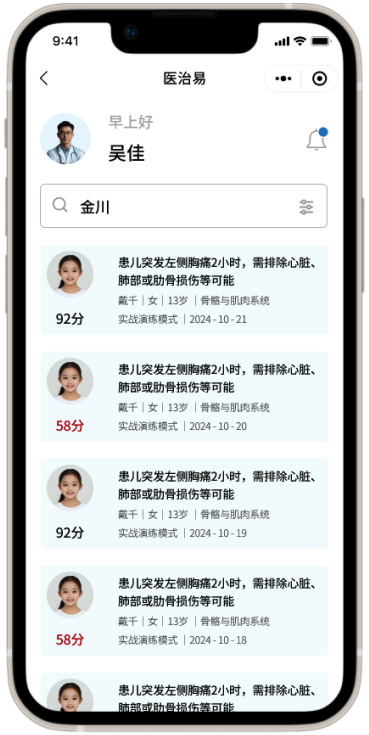
Zhang et al. Human or LLM as Standardized Patients? (arXiv:2511.14783) · Gao et al. “It Talks Like a Patient, But Feels Different” (CHI 2026).

Virtual Patients for Medical Education

- Train doctors in realistic hospital scenarios









In-house MedGen video generation
In-house TTS-LoRAHub speech synthesis
In-house HeadVid audio-driven digital humans

Application: AI Standardized Patients for Medical Training

Anhui University of Science and Technology

Simulate the whole consultation with AI to train clinical thinking: virtual patients + real-time feedback

Teaching value

For students:

- train diagnostic reasoning; fewer missed or wrong diagnoses

For teachers:

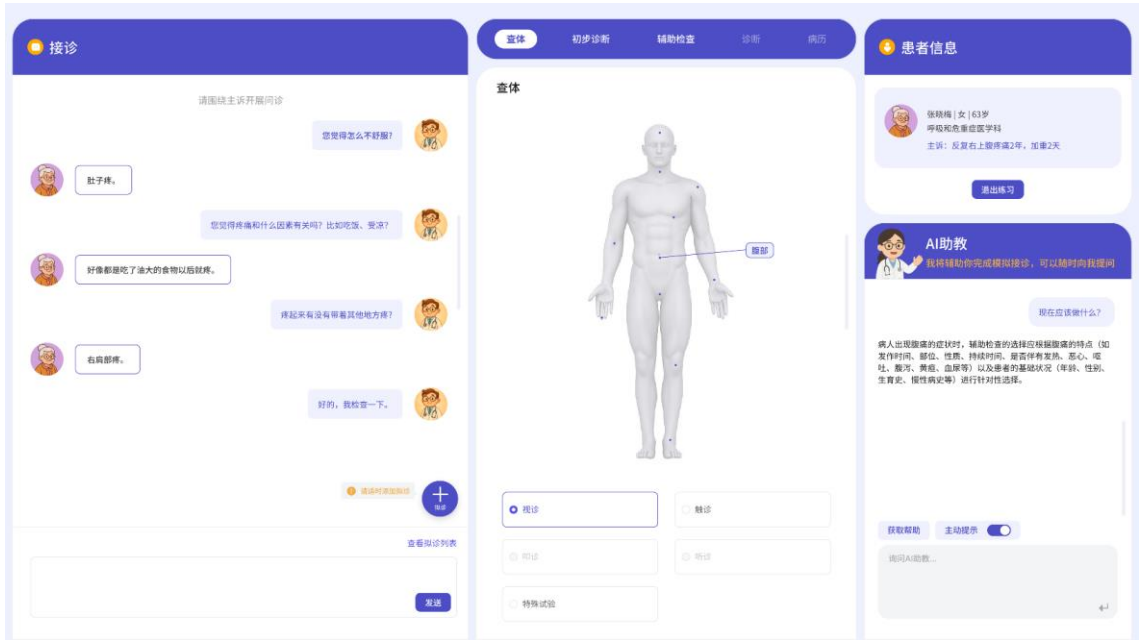
- lower standardized-patient costs; traceable, digital teaching

Goals

full-workflow feasibility

effective feedback

live demonstration

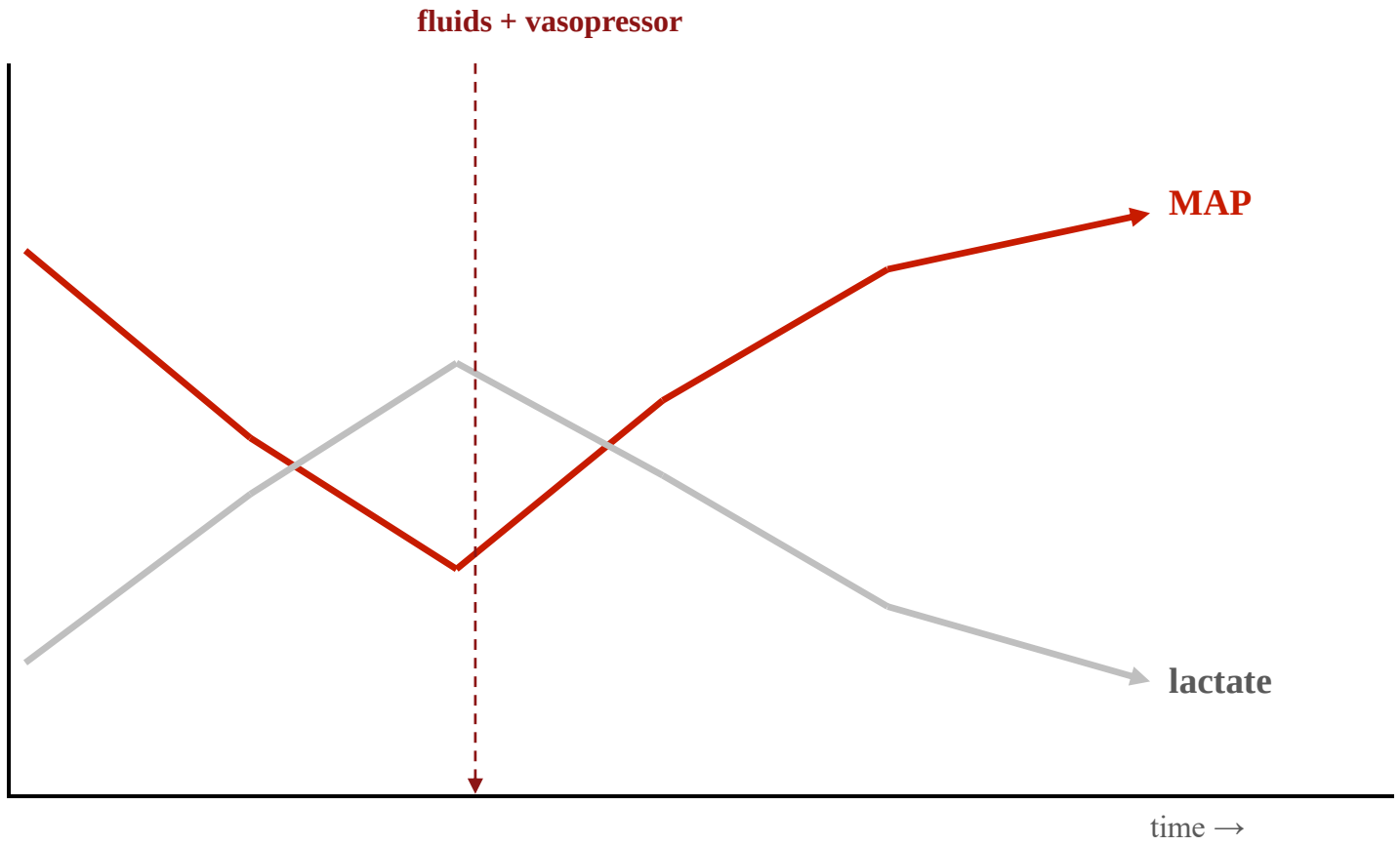


real-time correction **reasoning analysis** **personalized practice**

6.2 Patient Trajectories and Clinical World Models

A patient is not a static vignette: the body responds to every action. Models must learn how states evolve.

A Patient Is a Trajectory, Not a Vignette



What models must learn

- state → action → next state
- delayed, noisy effects of treatment
- Scale trajectories: ICU records (e.g. MIMIC), physiology simulators and learned generators

Scaling patients means scaling physiology over time — the data a sequential decision maker needs.

Illustrative septic-shock course · Johnson et al. MIMIC-IV (2023).

Clinical World Models

$s_{t+1} \sim W(s_t, a_t)$ — predict how the patient responds to an action

State s_t
vitals, labs, notes, images, history

Action a_t
orders: tests, drugs, fluids, procedures

Transition W
learned from EHR + physiology

Reward
outcomes: stabilization, survival, cost

Train
RL for clinical agents without risk to real patients

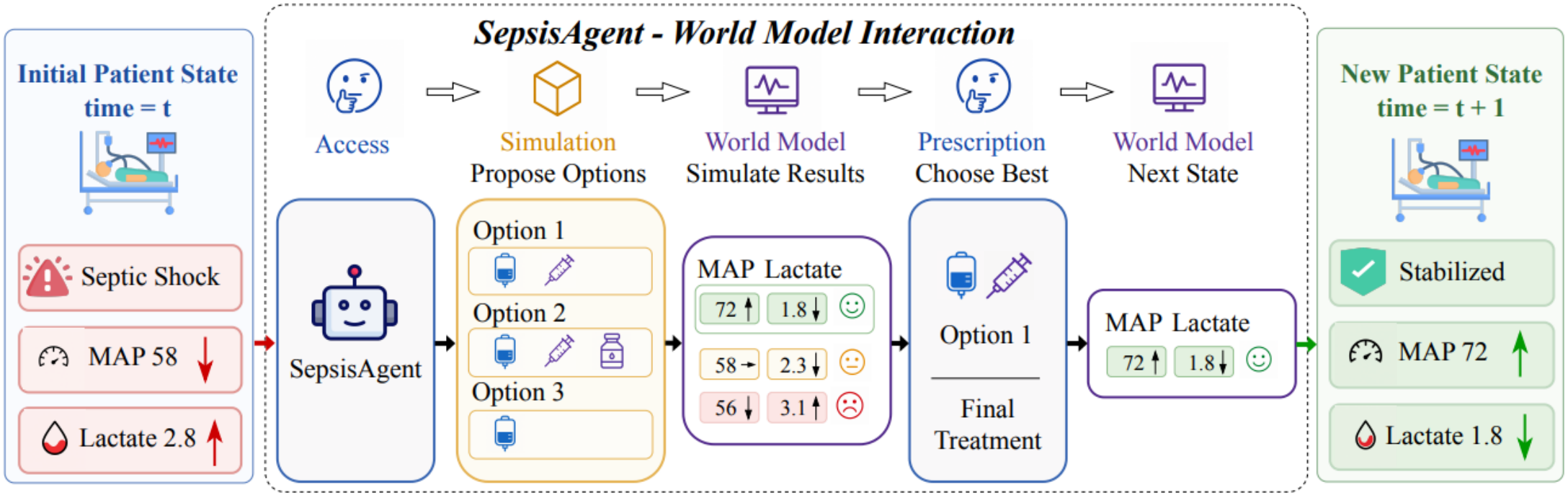
Ask "what if?"
counterfactual treatment comparisons

Evaluate
score agents on outcomes, not on answer keys

A world model turns medicine into an environment: the agent proposes, the world model answers.

Wu et al. Agentifying Patient Dynamics within LLMs through Interacting with Clinical World Model: A Case Study of Sepsis (2026).

Clinical World Model: Sepsis Case Study



Minghao Wu, Yuting Yan, Zhenyang Cai, Ke Ji, Chuangsen Fang, Ziying Sheng, Xidong Wang, Rongsheng Wang, Hejia Zhang, Shuang Li, Benyou Wang, Hongyuan Zha. Agentifying Patient Dynamics within LLMs through Interacting with Clinical World Model: A Case Study of Sepsis. Submitted to NeurIPS 2026.

6.3 The Agentic Hospital

Medicine happens in workflows. AI must enter them — first in simulated hospitals that speak the language of real ones.

The Agentic Hospital: Four Layers of Simulation

People

virtual patients and doctors; realistic audio–video (HealthTTS, TalkVid, MedGen)

Social interaction

dialogue, reports, orders and payments between all roles

Physical space

3D hospital from floor plans and drone video (Gaussian splatting)

World mechanisms

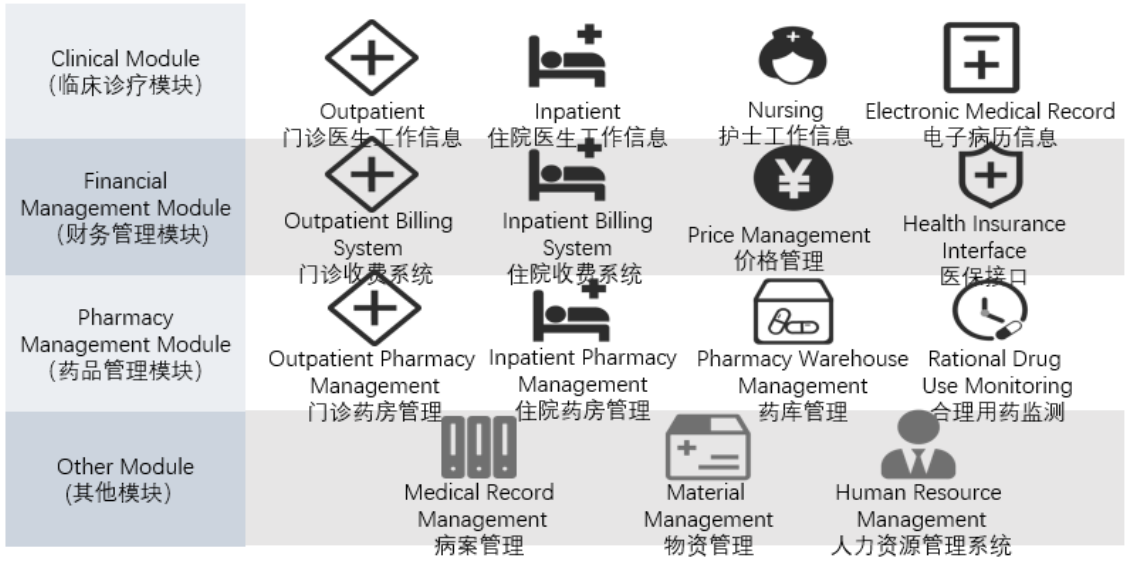
insurance, triage, scheduling, incentives of people and institutions

Bridge: a real Hospital Information System (HIS). Generate many hospitals, all speaking the HIS language — tools built in simulation plug into real hospitals.

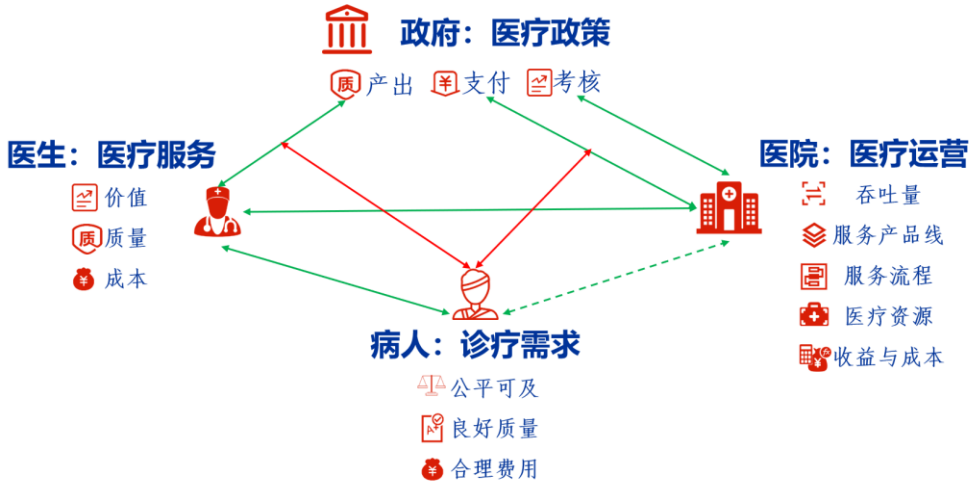
Related: Agent Hospital (Tsinghua, 2024) trains doctor agents on simulated patients (MedAgent-Zero).

Wang et al. (2026) · Chen et al. Digitalized Hospital (in preparation) · Li et al. Agent Hospital (arXiv:2405.02957).

Hospital Simulation: Who Pays in an Aging Society?



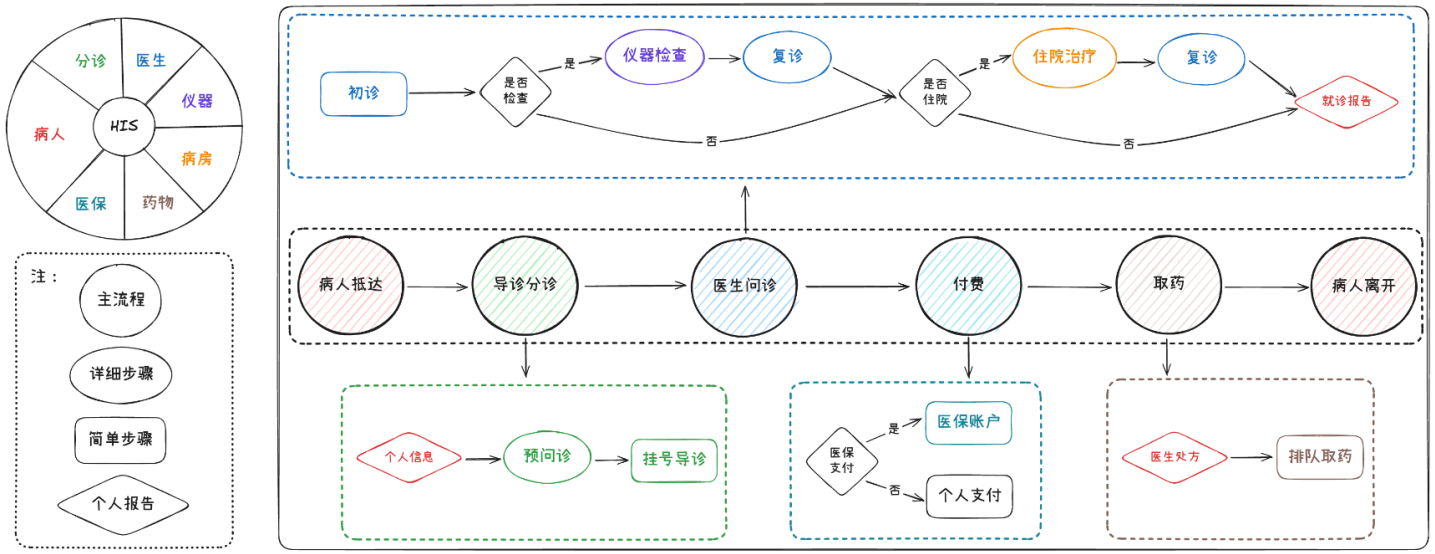
In an aging society, insurance payment is the hard pr



Why Connect to a Real Hospital Information System (HIS)?

Data move seamlessly between the virtual and the real hospital; any AI tool built in the virtual hospital plugs directly into the real system. CUHK-Shenzhen Hospital has deployed a real HIS on our university servers, with controlled sample data.

Digitalized Hospital: Simulating the Full Clinical Workflow

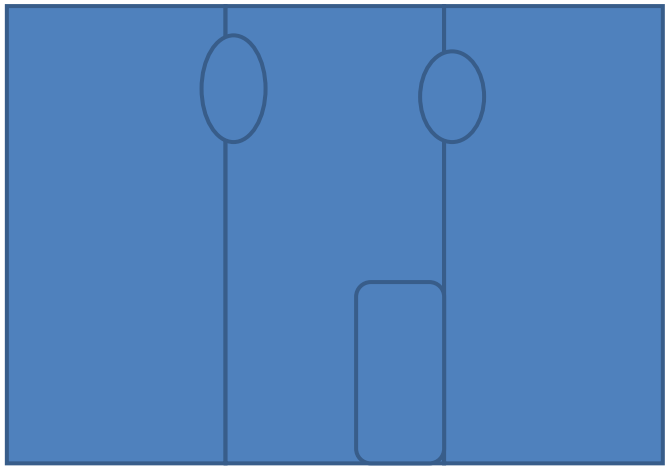


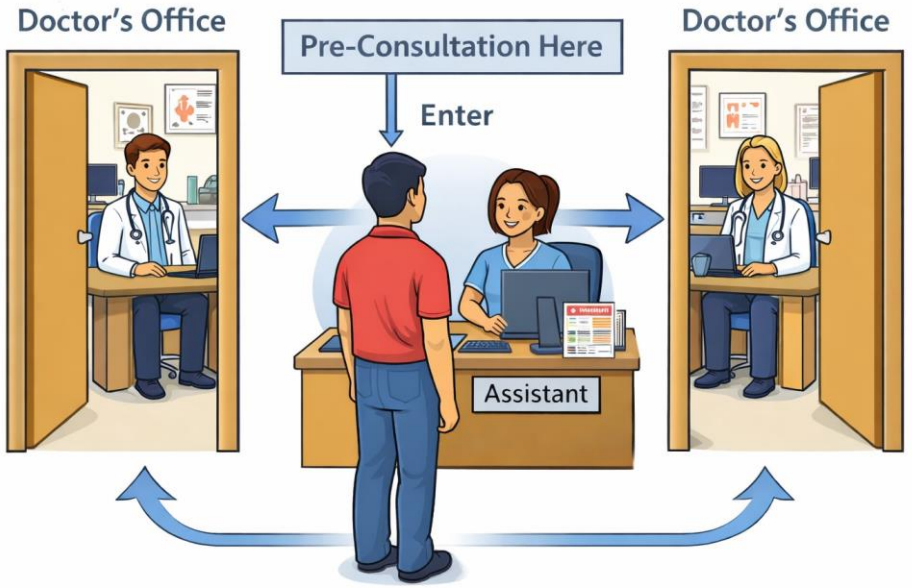
HuatuoGPT is connected to 12 public hospitals and ~200 community health centers in Longgang, deployed for pre-triage, pre-consultation and family-doctor services.
But hospital information systems are complex and heterogeneous, so AI is hard to embed. First harden full-workflow AI in a virtual hospital, then migrate it gradually to real hospitals.

Ziqi Chen, Jiale Han, Zhongtian Jiang, Xiang Li, Benyou Wang. Digitalized Hospital: Simulating Hospitals with real-world Health Information Systems. In preparation.

A Sandbox for Hospital Operations

- Is a human pre-consultation step worth it?





- What if we remove this assistant?
- Can AI replace it? With what trade-offs?
- How does it differ across departments?

Digitalized Hospitals as a Tool for Agents

Agents adjust and reflect on their operation and management strategies using feedback from the digitalized hospital



A Sandbox for AI Applications

Medical-record writing: rubric RL driven directly by record requirements, scored by the hospital's real quality-control system; needs access to all patient information (e.g., inpatient records) and evidence sources such as UpToDate; deployed on an upgraded HuatuoGPT.

6.4 Virtual Cells and Virtual Organs

Below the patient lies biology. Simulating cells and organs connects medical AI to drug discovery.

Virtual Cells: Simulating Biology for Drug Discovery

Cell state
single-cell expression

→

Perturbation
gene knockout, drug, dose

→

Virtual cell model
predicts the response

→

In-silico screen
rank targets and drugs

Where the field is (2025–2026)

- Arc Institute's State model: trained on ~170M observational and >100M perturbed cells
- Virtual Cell Challenge 2025: 5,000+ participants from 114 countries; 2026 edition is zero-shot
- Honest caveat: beating simple baselines is still hard

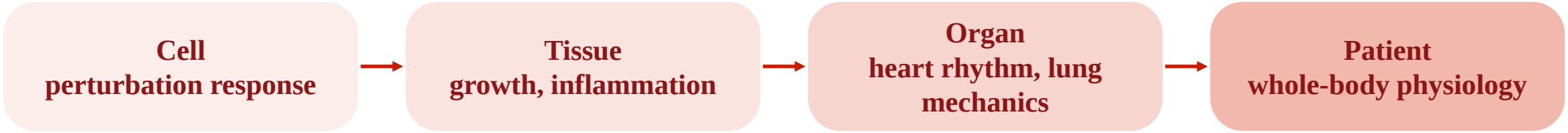
Our angle

- MicroVerse: generate micro-world videos (cells, DNA, alveoli) from text
- Language as the interface: ask a virtual cell questions, get explanations

If we can simulate how cells respond, drug discovery becomes search in an environment.

Arc Institute: State (2025), Virtual Cell Challenge 2025 wrap-up and 2026 · Wang et al. MicroVerse (ICLR 2026).

From Cells to Organs to Patients: Multi-Scale Digital Twins



Drug development

predict efficacy and toxicity before trials; in-silico trial arms

Personalized therapy

simulate a specific patient's organ before treating it

Training environments

agents learn treatment strategies on twins, not on people

Open question: one foundation model that speaks every scale — molecules, cells, organs and clinical language — or many coupled simulators?

Bunne et al. How to Build the Virtual Cell with AI (Cell, 2024) · literature on cardiac digital twins and in-silico trials.

6.5 Community-Based Evidence-Based Medicine

Evidence changes every day. A community of experts, helped by AI, can keep it alive and trustworthy.

Community-Based Evidence-Based Medicine



The evidence pyramid

AI drafts — from guidelines and literature

↓

Experts edit & sign — credit and incentives

↓

Users give feedback — inside agent apps

↓

Evidence API — retrieval for every medical agent

Market signal: OpenEvidence, an evidence-grounded assistant, reports use by ~65% of US physicians. Our goal: an open, AI-native UpToDate maintained by a community.

Sackett et al. Evidence-Based Medicine (BMJ 1996) · NBC News on OpenEvidence (May 2026; company figures).

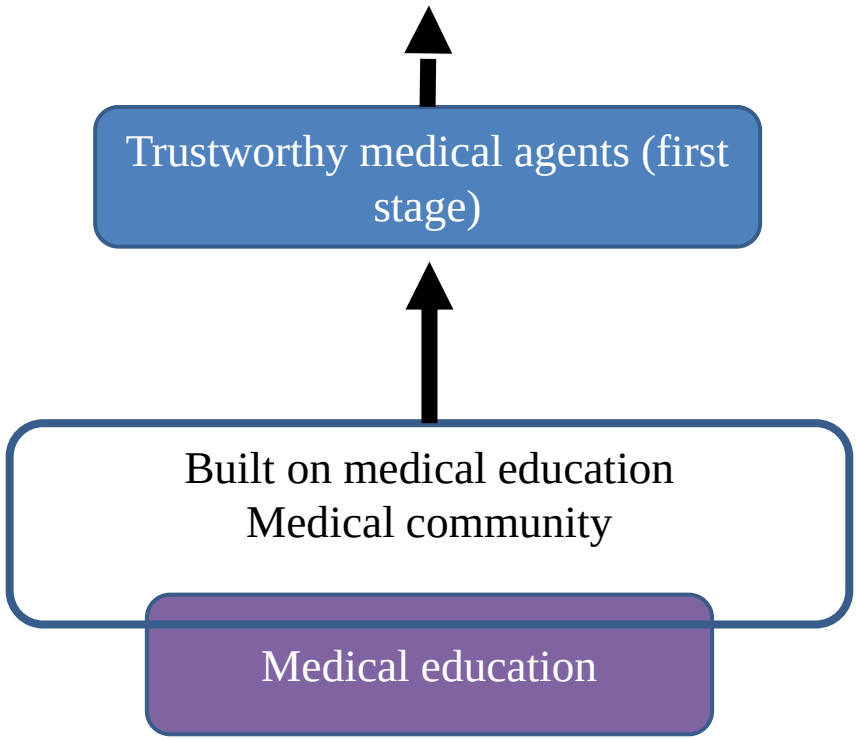
UpToDate: Today's Point-of-Care Reference



AI-Native UpToDate

AI drafts the core medical content: a complete, evidence-based body of knowledge (guidelines, Wikipedia). The community keeps it updated: experts are proud to edit (“hospital-certified pediatrician”), AI helps them, and incentives (e.g., free model API access) reward contributors. Related work KBAgent: AI + professional documents build general and specialist knowledge graphs, experts verify, user feedback refines. Exposed as an evidence API with a retriever.

Our Path: Education → Community → Trustworthy Agents → GMAI









Finally: generalist medical AI

Specialist models are fused into trustworthy agents, refined with the community

The community links AI industry, hospitals, doctors, students and the public to co-develop specialist models

AI makes medical students learn faster and trains the next generation of doctors to work with AI

Starting from medical education, we build an open, AI-friendly medical community; its real needs and feedback gradually yield trustworthy medical agents, and finally generalist medical AI.

Final Goal: Open, Community-Driven Medical Agents

LLMs

- Deepseek
- Qwen
- Baichuan
- **HuatuoGPT**
- **Specialized LLMs**
- O1-like LLMs
- Cost-effective LLMs
- Local LLMs

Leaderboard

- Agents
  - LLM routing
  - Agent Framework
  - Text Embedding
  - Confidence estimation
  - Trustworthy retrieval source

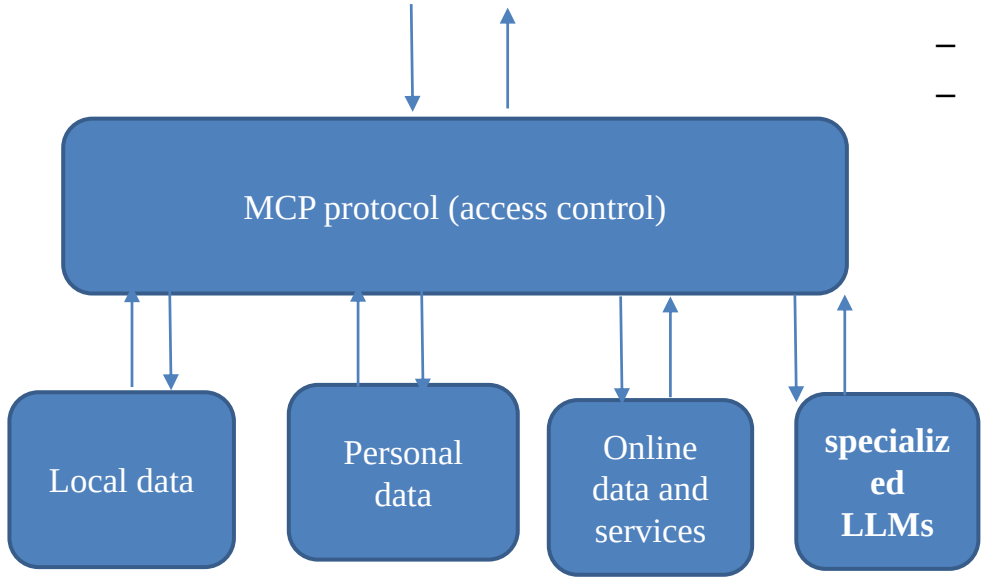
Need and feedback

Build a community

- Community
  - Users from
    - **Medical education**
    - **Science popularization**
    - AI guides/book/media for doctors
    - Online/offline seminars
    - Medical schools
  - Medical tool hub (MCP)
  - Corpora and dataset
  - Benchmarking
  - Collaborated on **specialized LLMs**

Cost-effectiveness

There is no single model which are always the best



6.6 Data, Evaluation and Infrastructure

Every environment above rests on data, evaluation and infrastructure that someone has to build.

Data and Infrastructure Underneath Everything

Data

trustworthy corpora (Med-Pajama), interleaved multimodal data, patient trajectories, rubrics

Compute & privacy

on-premise deployment, HIS sandboxes, privacy-preserving training

Tools

a medical tool hub over MCP; skills for clinical tasks

Evaluation

live, leakage-free benchmarks (LiveClin, Doctors' Last Exam) with rubrics

Community

doctors, students and engineers who label, edit and use

Governance

consent, audit trails, ethics review, liability

Unglamorous, but decisive: whoever builds this layer sets the standards for medical AI.

Evaluation as Infrastructure: Benchmark Engineering

Evaluation needs consensus and support from the leading players!

LiveClin: A Live Clinical Benchmark

Scenario: A 30-year-old African American woman presents with gradually worsening pain in her left knee that has been present for two years and has recently intensified. She denies trauma, fever, or systemic symptoms. Examination reveals focal tenderness over the proximal tibia with mild discomfort on range of motion; neurovascular status is intact and vital signs are normal.

Stage 1: Initial Assessment

Question: Anteroposterior radiography of the left knee is obtained (Figure 1 shows a plain X-ray). Which of the following is the most likely diagnosis suggested by the imaging findings?



Figure 1 (X-ray)

Stage 2: Diagnostic Work-up

Question: Magnetic resonance imaging of the knee is performed (Figure 2 shows MR sequences), and a staging chest radiograph is unremarkable (Figure 3). Which of the following is the most appropriate next diagnostic step to establish a definitive tissue diagnosis?

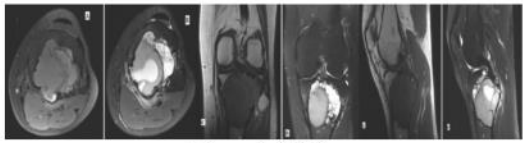


Figure 2 (MRI)

Stage 3: Pathologic Diagnosis

Question: Histologic sections from the biopsy are shown in Figure 4 (hematoxylin-eosin stain). Based on the microscopic appearance, which of the following best characterizes the lesion?

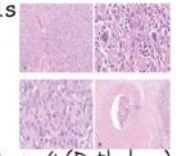


Figure 4 (Pathology)

Stage 4: Initial Therapeutic Planning

Question: Following confirmation of a high-grade malignant primary bone tumor, the oncology team plans neoadjuvant systemic therapy. Which regimen is considered standard first-line chemotherapy for this type of high-grade bone sarcoma?

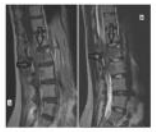


Figure 5 (MRI)

Stage 5: Disease Progression

Question: Twenty months after the initial presentation of a metastatic malignant giant cell tumor (GCT) involving the spine, the patient reports severe lumbar pain radiating to the right leg. MRI of the lumbar spine is obtained (Figure 5 shows sagittal T1- and T2-weighted images). Which of the following complications is most consistent with the findings on this MRI?",



Figure 3 (X-ray)



Figure 6 (CT)

Stage 6: Management of Refractory Metastatic Disease

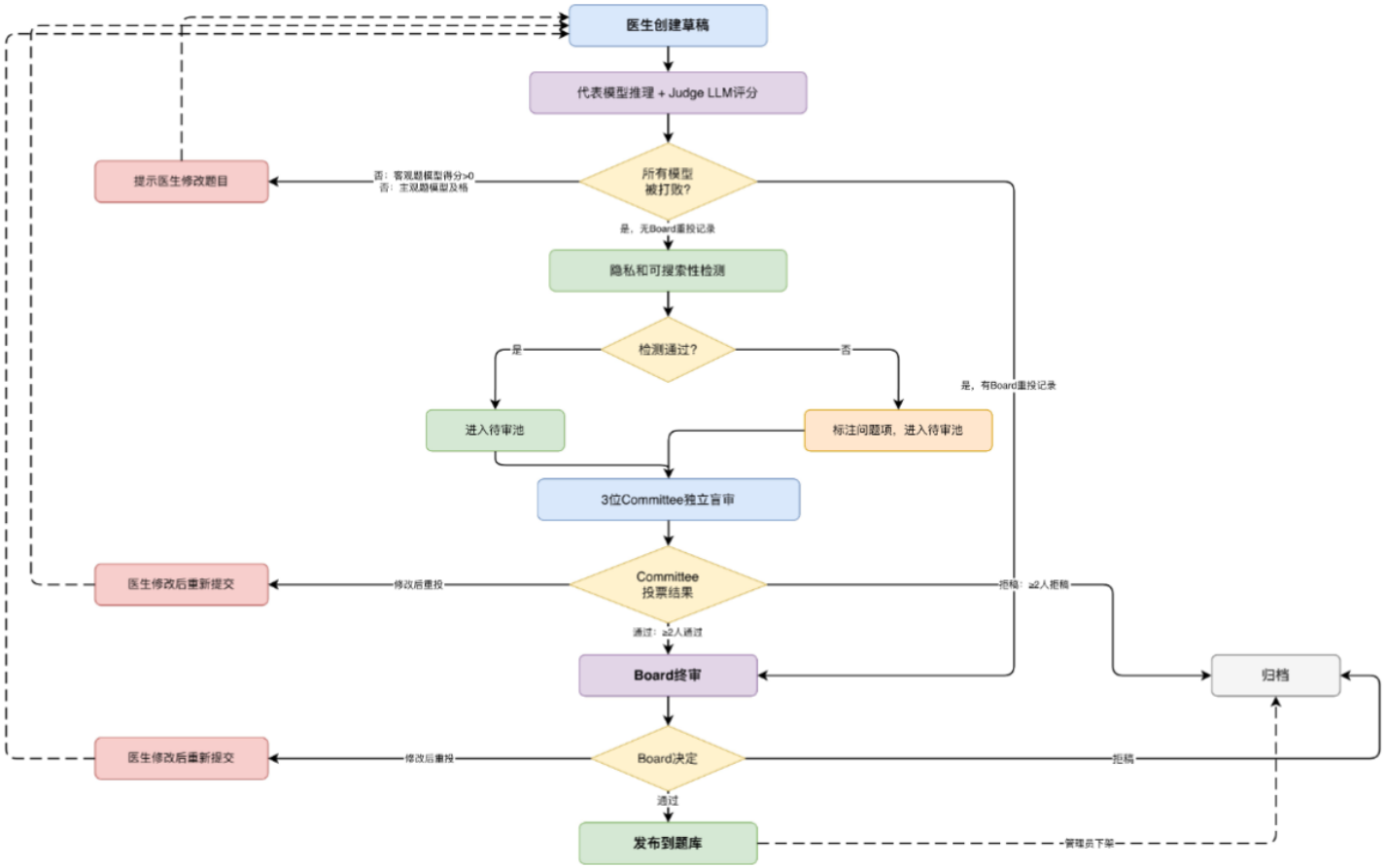
Question: Repeat thoracic imaging now demonstrates numerous bilateral pulmonary nodules (Figure 6 shows a non-contrast CT of the chest). The patient progressed despite MAP and cisplatin/doxorubicin regimens. Which of the following systemic is the most appropriate next-line therapy for her refractory metastatic disease?

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, Benyou Wang. LiveClin: A Live Clinical Benchmark without Leakage. ICLR 2026

Doctors' Last Exam: Design

- **Updates: rolling, monthly or quarterly, with a white paper (see LiveClin, with Ant Healthcare).**
- **Leaderboards: board A releases questions but not answers; board B hides both (models are submitted for testing).**
- **Production: experts write questions; a parallel track synthesizes them with agents (eventually replacing manual writing). Every item is e-signed by an expert.**
- **Format: rubric-based grading in the spirit of HealthBench (rubrics preferred, not mandatory).**
- **Partnership: co-built with industry partners (e.g., Ubiquant, Baichuan, Ant Healthcare, SenseTime Healthcare).**
- **Quality over quantity: every item has a named expert author; a committee reviews items and rewards good ones.**

Dentists' Last Exam



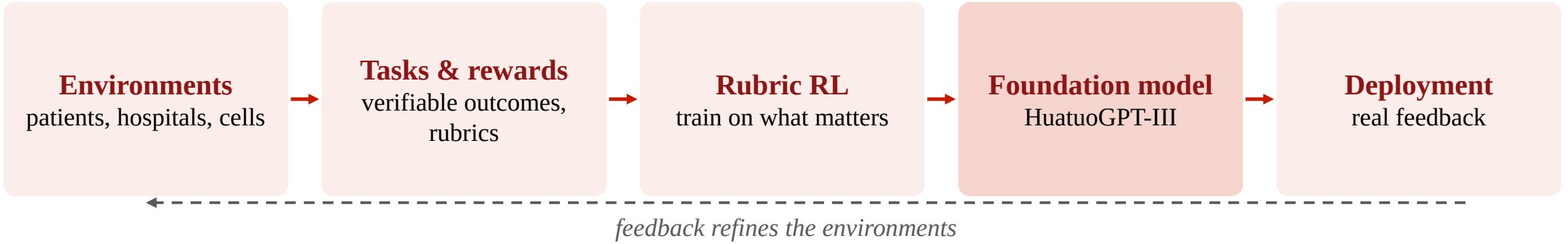
Evaluation Through Agent Routing

Agent applications need reputation and traffic; a single developer video can make a product spread. Measure real effectiveness inside agent applications, and publish a quarterly technical report (like OpenRouter's).

6.7 Back to Better Foundation Models

Environments are not the end: they produce tasks and rewards that train the next, better foundation model.

From Environments Back to Better Foundation Models



The model remains the core asset — but it is now trained by environments rather than only by text.

HuatuoGPT-III: RL-only domain adaptation from a base model, with trustworthy mid-training (Med-Pajama) and rubric RL.

Better Foundation Models: HuatuoGPT-III

- Rubric RL, with multimodal inputs and tool calls (images, audio, …)
- Mid-training with trustworthy corpora (Med-Pajama)
- Evidence-based RAG
- Reuse existing small vision models and rich interleaved multimodal data

Submission and Formatting Instructions for ICML 2026

Model	Open-End(HealthBench)			Close-End		
	Total	Hard	Cons.	Medbullets(5ops)	MMLU-Pro(Med)	MedXpertQA(Text)
Leading Proprietary Models						
GPT-5.2 (high)	**59.9**	**40.5**	**65.0**	88.9	90.0	54.2
GPT-5-Chat	51.9	23.4	56.2	81.5	86.8	45.3
DeepSeek-V3.2-Thinking	55.1	23.0	57.8	83.5	88.1	44.3
Gemini-3-Pro	47.9	18.5	54.5	**90.5**	**93.0**	**71.3**
Our Experiments (Base Model: Qwen3-8B-Base)						
Qwen3-8B	45.9	10.9	50.9	56.6	78.5	18.8
Qwen3-8B-Base	22.1	0.0	23.7	30.7	41.2	11.7
No Teacher						
w/ Pure RL	59.8	25.2	61.4	48.1	75.0	20.0
Teacher: GPT-5-Chat (Non-Thinking)						
w/ SFT+RL	63.6	37.3	65.0	61.9	78.2	21.5
w/ OnePO(Ours)	65.4(+43.3)	39.1(+39.1)	66.8(+43.1)	64.0(+33.3)	81.2(+40.0)	24.9(+13.2)
Teacher: DeepSeek-V3.2-Thinking						
w/ SFT+RL	64.5	40.7	65.4	63.5	80.0	22.5
w/ OnePO(Ours)	**67.2**(+45.1)	**44.5**(+44.5)	**68.0**(+44.3)	**65.2**(+34.5)	**82.0**(+40.8)	**25.9**(+14.2)

Xidong Wang, Shiwei Lyu, Hao Zhu, Lei Liu, Chaohe Zhang, Jian Wang, Jinjie Gu, Benyou Wang, Yue Shen. From Rubrics to Principles: Scalable Healthcare Alignment with Clinician-Grounded Principles.
Junying Chen, Xinyuan Xie, Ziniu Li, Benyou Wang. OnePO: Direct One-stage Policy Optimization for SFT-free Domain Adaptation.

HuatuoGPT-III: RL-Only Domain Adaptation

Submission and Formatting Instructions for ICML 2026

Model	Open-End(HealthBench)		Close-End		
	Total	Hard	Medbullets(5ops)	MMLU-Pro(Med)	MedXpertQA(Text)
Representative Baselines					
GPT-5-Chat	51.9	23.4	81.5	86.8	45.3
Gemini-3.1-Pro-Preview	54.5	21.5	**89.9**	88.6	**72.0**
GPT-5.2 (high)	**59.9**	**40.5**	88.9	**90.0**	54.2
Qwen3-8B	45.9	10.9	56.6	78.5	18.8
Qwen3-32B	48.2	15.3	71.8	82.9	25.8
DeepSeek-V3.2-Thinking	55.1	23.0	83.5	88.1	44.3
Our Experiments (Base Model: Qwen3-8B-Base)					
Qwen3-8B-Base	22.1	0.0	30.7	41.2	11.7
No Off-policy					
w/ Pure RL	59.8	25.2	48.1	75.0	20.0
Off-policy: GPT-5-Chat (Non-Thinking)					
w/ SFT+RL	63.6	37.3	61.9	78.2	21.5
w/ SeedRL(Ours)	65.4(+5.6)	39.1(+13.9)	64.0(+15.9)	81.2(+6.2)	24.9(+4.9)
Off-policy: DeepSeek-V3.2-Thinking					
w/ SFT+RL	64.5	40.7	63.5	80.0	22.5
w/ SeedRL(Ours)	**67.2**(+7.4)	**44.5**(+19.3)	**65.2**(+17.1)	**82.0**(+7.0)	**25.9**(+5.9)
Scaled HuatuoGPT-3 Models (Off-policy Seeded)					
HuatuoGPT-3-8B	67.5	45.4	66.2	82.3	24.4
HuatuoGPT-3-32B	70.3	46.1	76.6	86.0	28.7

Junying Chen, Xinyuan Xie, et.al. HuatuoGPT-3: RL-Only Domain Adaptation from Base Models via Off-Policy Seeding.

HuatuoGPT-III Roadmap

- **HuatuoGPT-III: mid-training for evidence-based medicine**
  - Built Med-Pajama, a high-quality trustworthy medical corpus
  - Large-scale mid-training plus retrieval to reduce hallucination and give trustworthy answers

- HuatuoGPT-Vision-II: a multimodal medical foundation model
  - Distill many existing expert vision models
  - Learn from large interleaved multimodal data (PubMed)

- HuatuoGPT-**O**3
  - Stronger multimodal reasoning: read case images like a doctor (zoom in/out, pan)

- Audio-driven
  - HuatuoGPT-Audio voice assistant
  - MedASR for real-time medical-record writing
  - HealthTTS for virtual patients

7 Applications and Ecosystem

Research matters when it reaches patients, doctors and students — and when it is shared openly.

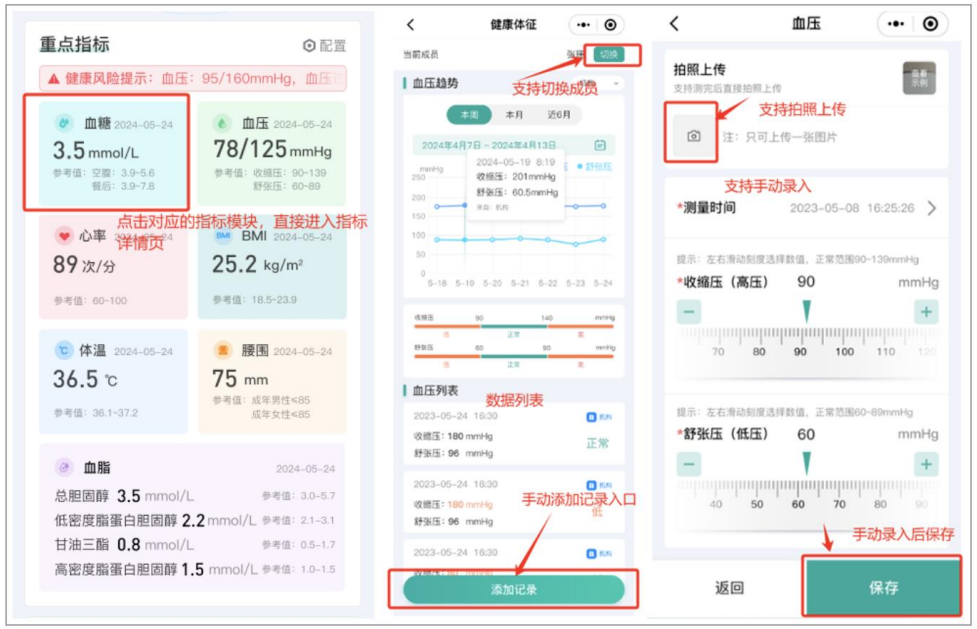
Deployment 1: Regional Triage Platform in Longgang



- **Coverage: 5 million people. HuatuoGPT passed public testing by the Shenzhen Health Commission in June 2023 and was deployed in 11 hospitals in Longgang — every district public hospital.**

- **Accuracy: 95.4%. China's first region-wide multi-hospital LLM deployment: 300K users, 520K+ interactions, 95.4% triage accuracy.**

Deployment 2: AI Family Doctor in Longgang



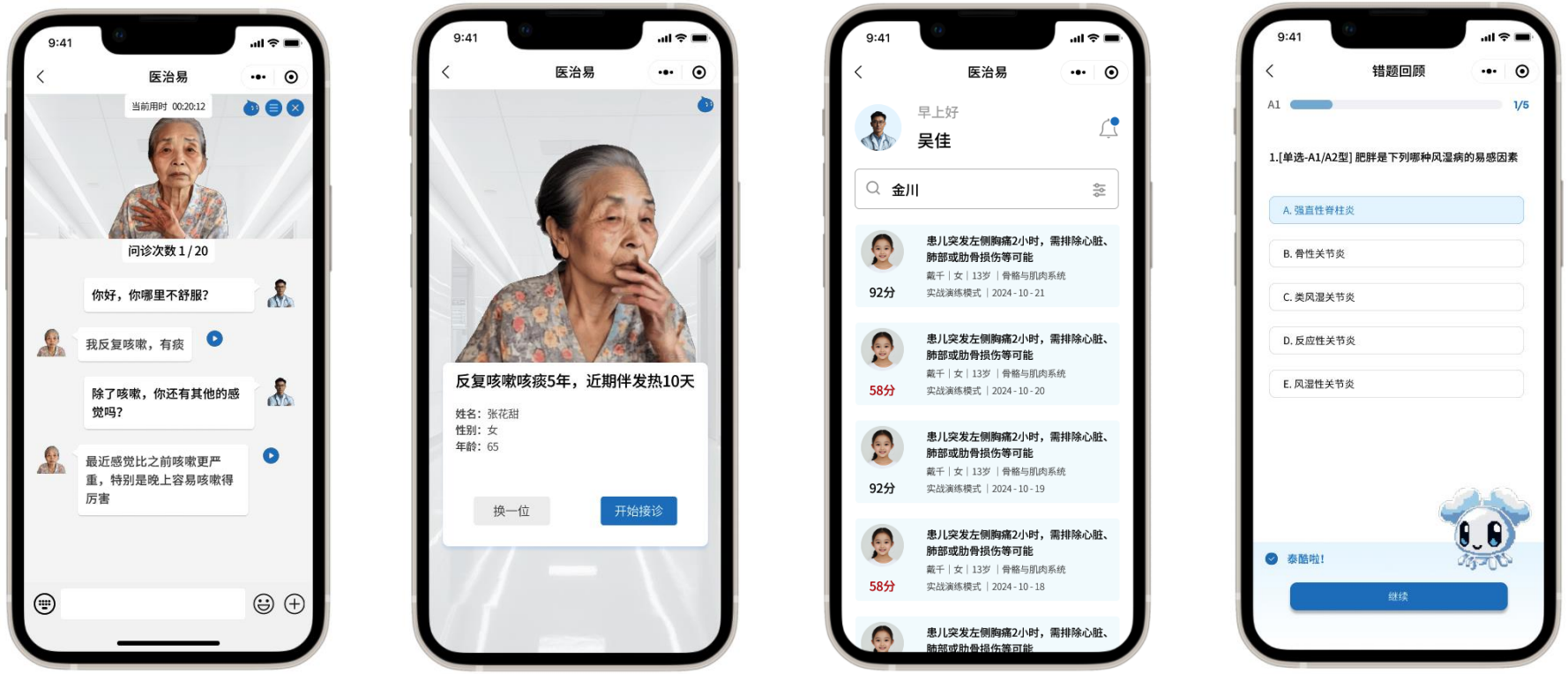


186 community hospitals and 12 local hospitals

it serves 5 million people

The Longgang AI family-doctor mini-program, built on HuatuoGPT, offers health queries, AI consultation and report interpretation; deployed in 12 hospitals and 183 community health centers.

Deployment 3: Simulated Patients for Training Doctors



Synthesize many AI-generated simulated patients from real cases to train doctors.

Deployment 4: Health Education with Video Generation



Generate health-education videos with video-generation models.

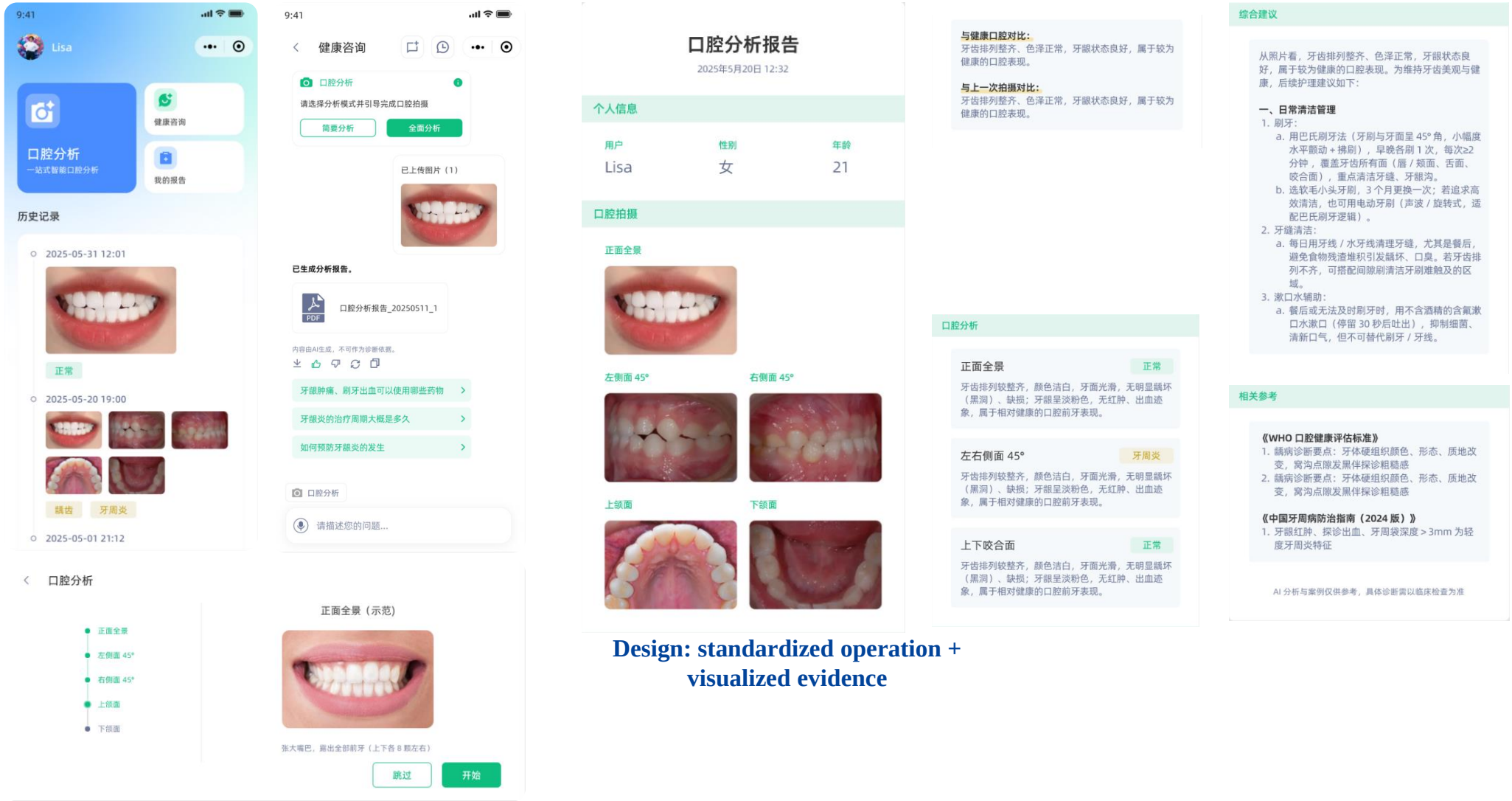
Junying Chen, Rongsheng Wang, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, Benyou Wang. Unlocking Medical Video Generation by Scaling Granularly-annotated Medical Videos.

Deployment 5: Personal Health Assistant



On-device consultation in the future — private data never leave the phone.

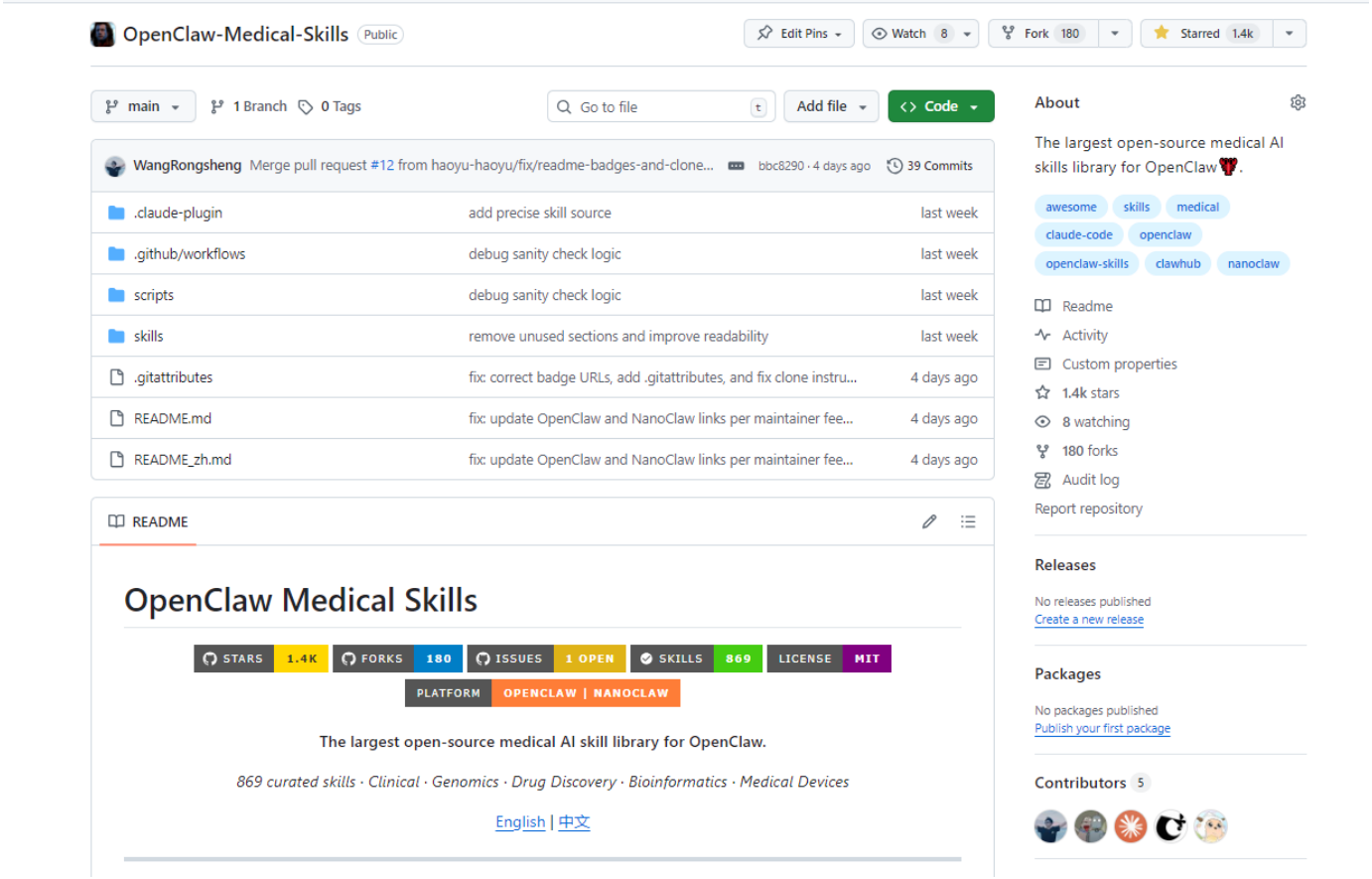
Deployment 6: Specialist Product — Dental Analysis



Design: standardized operation + visualized evidence

- Workflow: simple step-by-step guidance with real-time quality feedback
- Report: multi-dimensional analysis, predictions and personalized advice, with cited sources for trust

Medical Skills for Agents

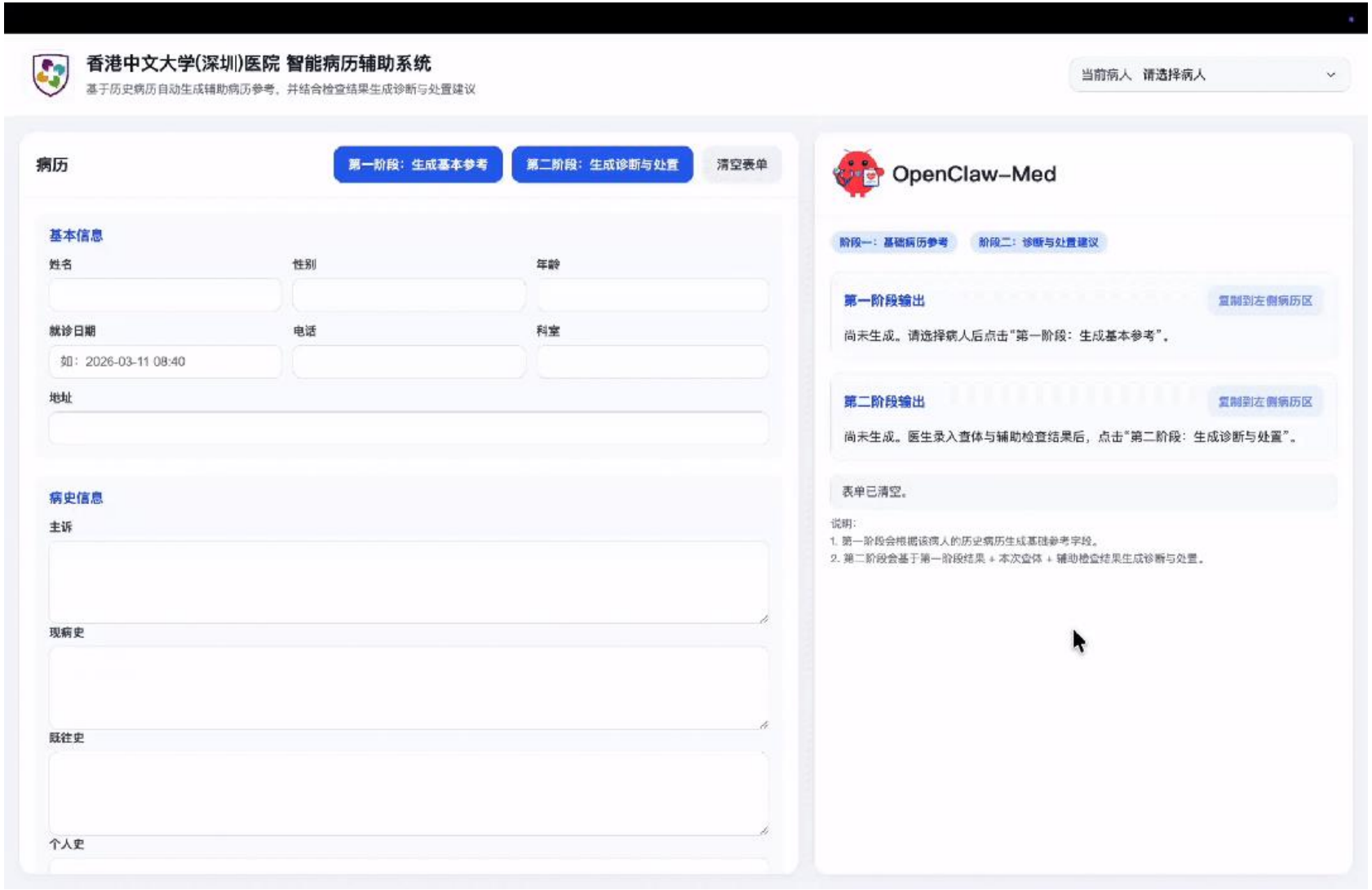


OpenClaw at CUHK-Shenzhen Hospital (1)





OpenClaw at CUHK-Shenzhen Hospital (2)



Medical Education: Happy-Medical-AI

- **The Happy-Medical-AI project**
  - https://github.com/FreedomIntelligence/HappyMedicalAI
  - AI training for health professionals: doctors who collaborate with AI
- Doctor education with patient twins
  - Simulated patients train doctors [1,2]; audio–video tech in fast development (lip-sync, digital humans, HealthTTS)
  - **Full-modality patient synthesis — the best training environment for AI and human doctors**
- Public health education
  - Video generation (MedGen)
  - Illustration generation
  - Micro-world simulator

- Start AI education early — it pays off for generations

[1] Zhiqi Gao, Guo Zhu, Huarui Luo, Dongyijie Primo Pan, Haoming Tang, Bingquan Zhang, Jiahuan Pei, Jie Li, Benyou Wang. " It Talks Like a Patient, But Feels Different": Co-Designing AI Standardized Patients with Medical Learners. ACM CHI 2026.
[2] Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Lei Zhou, Qianqian Xie, Benyou Wang. Human or LLM as Standardized Patients? A Comparative Study for Medical Education. arXiv:2511.14783.

Health Education in the Community

- **The Happy-Medical-AI project**
  - https://github.com/FreedomIntelligence/HappyMedicalAI
  - AI training for health professionals (online and offline)
- Start early with medical education
  - Audio–video patient twins under fast development
- Public health education
- AI training courses for doctors

中央美術學院
Central Academy of Fine Arts





- Start AI education early — it pays off for generations

Health Education: Dental Science (1)



Health Education: Dental Science (2)



Released with Shenzhen Stomatology Hospital (Pingshan), Southern Medical University

Startups from Our Team

自由动脉 FREEDOM AI

Jianquan Li founded Freedom AI (medical education and medical AI)



薯塔AI

Yang Xiao's Shuta AI: 1,000+ daily users, 60% coverage at CUHK-Shenzhen

Cognita

PhD graduate Zhihong Chen, after a Stanford postdoc, founded Cognita with Silicon Valley funding



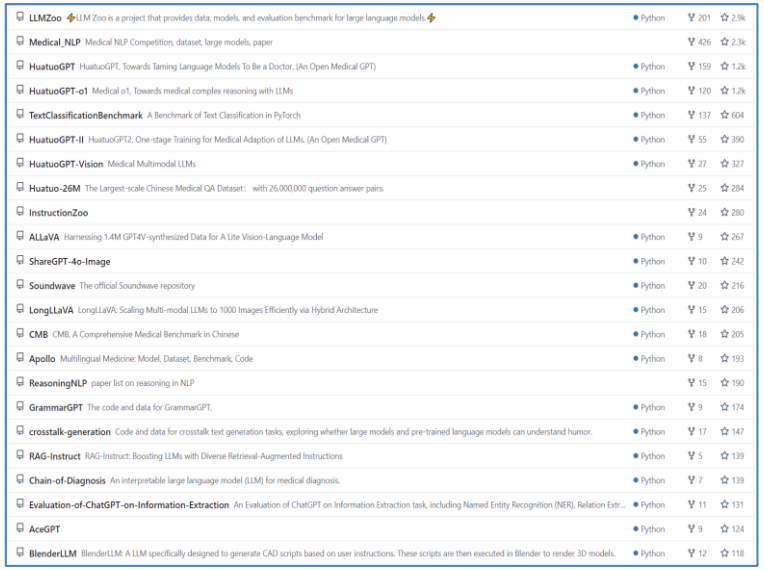
Yuxuan Zhu's TennSync app: 1,000+ users, 5,000+ hours saved for tennis players

METASTONE 是石科技

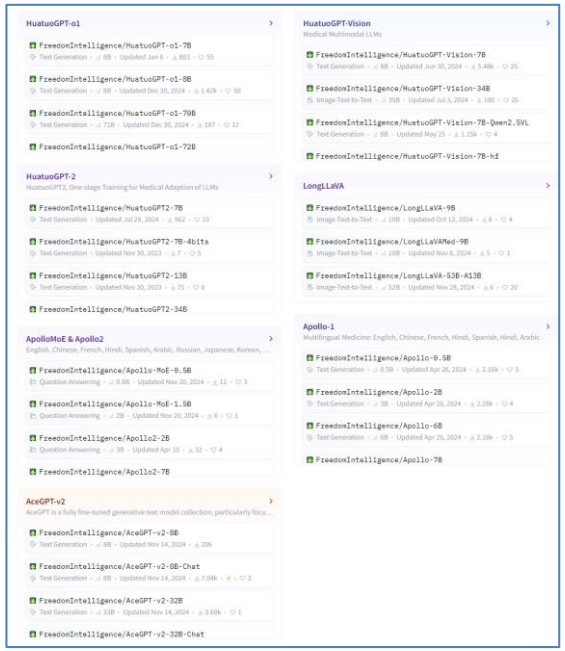
Jianye Hou's Metastone provides HPC and AI solutions, valued in the billions (RMB)

We Open-Source Everything: FreedomAI

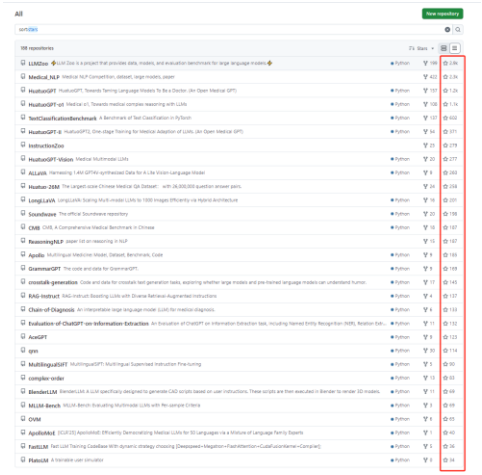
1.2k+ GitHub stars — we open-source whatever we can
https://github.com/orgs/FreedomIntelligence



1.3 million+ model and dataset downloads on Hugging Face
Ranked 14th among universities worldwide (Hugging Face)

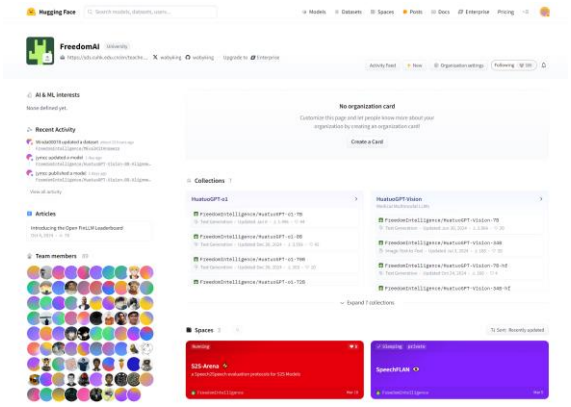


Sharing Our Research with the Community



10,000+ GitHub stars

https://github.com/FreedomIntelligence/

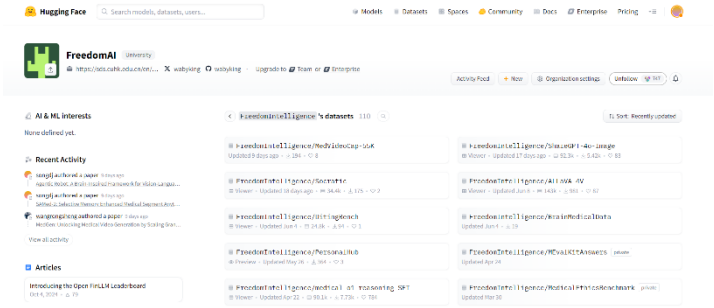


1M+ model downloads

https://huggingface.co/FreedomIntelligence/models



CMB leaderboard used by many companies



1M+ dataset downloads

https://cmedbenchmark.llmzoo.com/static/leaderboard.html

https://huggingface.co/FreedomIntelligence/datasets

8 Outlook

What medical AI needs next — and how to build it in the AI era.

Outlook: Five Bets for Medical AI

1. **Environments over datasets**
   the next gains come from places where AI can act and get feedback

2. **Trust before generality**
   earn deployment one specialty at a time

3. **Workflows over chatbots**
   value appears when AI enters triage, records, orders and payments

4. **Multi-scale simulation**
   from cells to hospitals — the substrate for drug discovery and clinical training

5. **Open communities**
   open models, open evidence, open benchmarks — shared by doctors and engineers

AI Entrepreneurship

We encourage undergraduates to build consumer medical applications for patients, doctors or medical students — backed by our technical team, SLAI and the CUHK-Shenzhen School of Medicine, with hardware (NVIDIA DGX Spark), technical guidance and small seed funds — and to connect specialties for social impact.

Organization: Smaller, Denser, More Decentralized

- AI-native startups are shifting from traditional company structures toward extremely lean teams, high-talent-density organizations, and community-driven coordination.
  - **Radically smaller organizations:** from traditional companies to "super individuals"; the rise of the **one-person company** (OPC)
  - **Native global collaboration:** open-source communities become the earliest organizational form; enthusiasts join organically, and GitHub acts as a coordination layer
  - **Work, users, and community come first: products, users, and communities are often built before formal company structures emerge.**
  - **Talent density matters more than full specialization:** teams emphasize highly capable, cross-functional people rather than rigid division of labor (Everyone is full-stack)
  - **Bottom-up innovation becomes stronger:** outstanding, ambitious young people can initiate innovation directly, while larger organizations often suffer from communication overhead
- **Core idea:**
  The basic unit of entrepreneurship is shifting from the large organization to the small team, the super individual, and the globally coordinated community.

Example: A One-Person Telehealth Company



	Medvi	Hims & Hers
2025 Revenue	$401M	$2.4B
2026 Run Rate	$1.8B	N/A
Net Margin	16.2%	5.5%
Employees	2	2,442
Outside Funding	$0	Public (NYSE: HIMS)

Matthew Gallagher (41) used $20K and two months to build a GLP-1 telehealth company from his living room in Los Angeles $401M revenue in year one, heading to $1.8B this year — one person plus a stack of APIs.

https://www.forbes.com/sites/josipamajic/2026/04/02/ai-and-20000-helped-one-man-build-a-18-billion-telehealth-startup

Product and Growth: Product as Marketing, Building in Public

- In the AI era, product, distribution, marketing, and founder presence are increasingly fused into a single system.
  - **Product as marketing:** growth is embedded in the product itself; great products advertise themselves, especially in consumer markets, enabling breakout adoption, exponential growth, and social virality
  - **Engineers go directly to market:** technical builders must also think like product people
  - **Founder as IP:** the founder often becomes part of the product and part of the distribution strategy
  - **Building in public becomes the default:** startups increasingly share iterations, roadmaps, feedback, and even problems openly
  - **Transparency becomes a competitive advantage:** public building is not only marketing, but also a way to attract users, developers, trust, and community participation
- **Core idea:**
  Growth, product design, distribution, and founder identity are no longer separate functions — they increasingly reinforce one another.

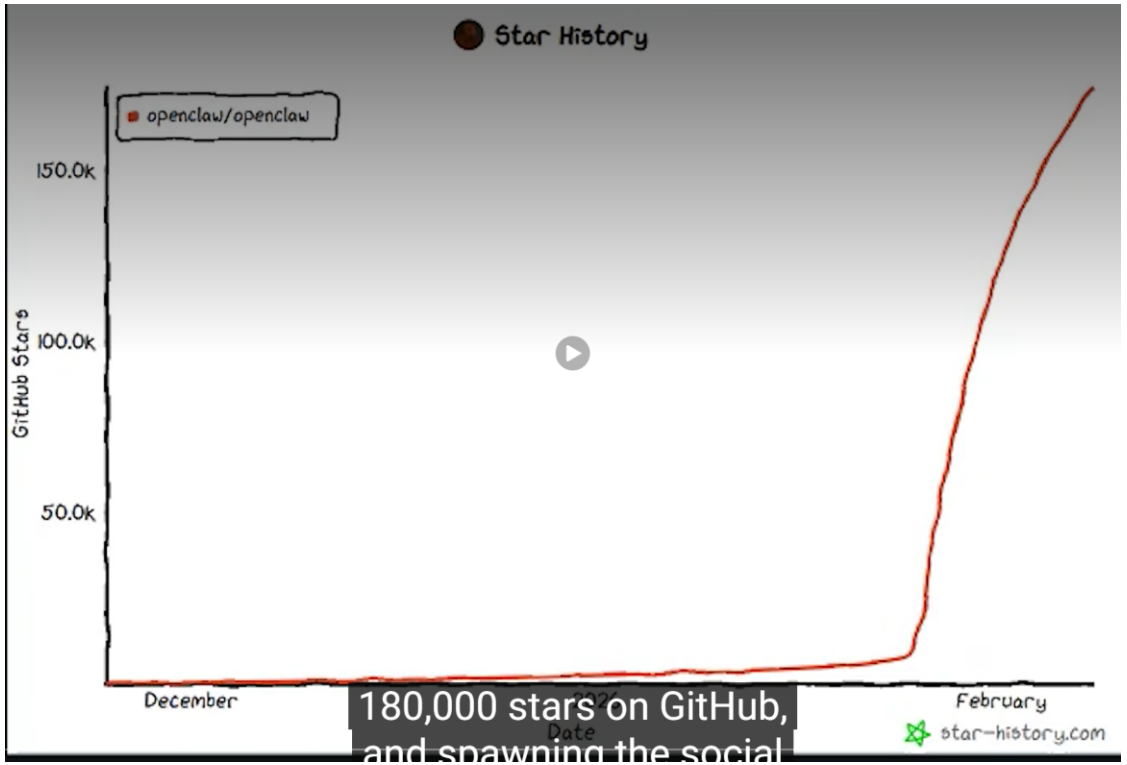
Speed and Iteration

- AI dramatically compresses the cost of experimentation, making speed and adaptability central to survival.
  - **The pace is much faster:** rapid, agile development becomes a basic requirement
  - **Experimentation is cheaper:** teams can test ideas, pivot, and switch directions more easily
  - **Business models are validated faster, but also copied faster:** AI lowers both the cost of building and the cost of imitation
  - **The winners are often not the most perfect, but the fastest to evolve:** long-term success comes from continuous learning and adjustment, not from getting everything right on day one
- **Core idea:**
  In the AI era, the key advantage is not just being early — it is being able to evolve continuously.

Example: Peter Steinberger and OpenClaw



Peter Steinberger



Leverage and Competition

- AI creates unprecedented leverage for individuals and small teams, but it also compresses the duration of competitive advantage.
  - **Extreme leverage:** *one person's productivity can equal that of 100 to 10,000* people in the past
  - **Advantages scale faster, but decay faster too:** products can spread quickly, but imitation also happens quickly
  - **Being first is no longer enough:** startups must compete on iteration speed, user relationships, brand recognition, and feedback-loop quality
  - **The moat shifts from static advantage to dynamic advantage:** durable edges increasingly come from how fast a company learns and compounds
- **Core idea:**
  AI amplifies both creation and competition. The new edge is not just capability, but the speed at which capability compounds.

Product Paradigm: From Static Software to Living Systems

- AI products are increasingly dynamic systems shaped by models, environments, users, and feedback loops.
  - **Products look more like living systems than static software**
  - **External model improvements can continuously enhance company capabilities**
  - **User and data loops matter more than feature shipping:** sustainable feedback loops are more valuable than isolated functionality
  - **Product boundaries become blurrier:** cross-domain fusion becomes more natural and more powerful
  - **User expectations are extremely high, while tolerance is very low**
- **Core idea:**
  AI products are no longer one-time software artifacts; they are evolving systems that improve through interaction, data, and iteration.

Founders: From Managing Resources to Amplifying Cognition

- The founder's role is shifting from coordination and resource control toward direction-finding, judgment, and human-AI orchestration.
  - **The founder's core capability shifts from resource management to amplified cognition**
  - **The most important skills become:** identifying direction, framing problems, defining products quickly, and organizing human-AI collaboration
  - **AI becomes a force multiplier for judgment and execution**
  - **Capital efficiency matters more:** fundraising is no longer the only central narrative, because AI reduces startup costs
  - **Global markets become even more important, especially under geopolitical fragmentation**
- **Core idea:**
  The future founder is less a manager of resources and more an amplifier of judgment, creativity, and execution through AI.

Join Us: PhD in AI Entrepreneurship at SLAI



Join SLAI as a PhD in AI Entrepreneurship (2026 Fall or 2027 Fall)

Takeaways

Why	Medicine is high-value and high-stakes: the right domain for serious AI.
Story	HuatuoGPT → HuatuoGPT-II → o1 → III: data, knowledge, reasoning, RL.
Goal	Generalist Medical AI is the goal; trustworthy medical AI is the necessary step.
Engine	Build environments: standardized patients, trajectories, world models, agentic hospitals, virtual cells, community evidence.
You	Every layer needs data, evaluation and infrastructure — and people who care about both AI and patients.

Thanks

wangbenyou@cuhk.edu.cn