Chiron-o1: Igniting Multimodal Large Language Models towards Generalizable Medical Reasoning

FUDAN UNIVERSITY 1905

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Presenter: Haoran Sun

Email: manglu3935@126.com

Motivation & Challenges

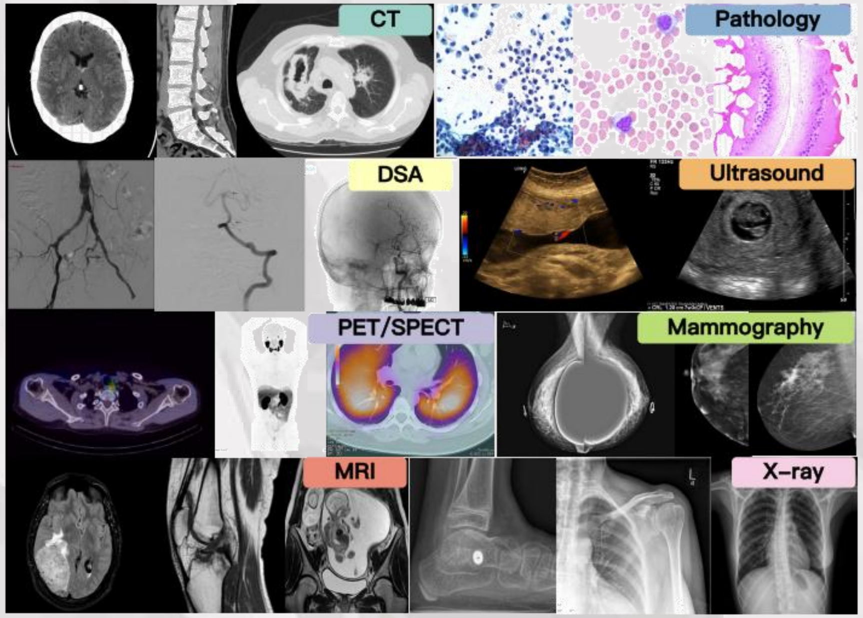
- Existing medical MLLMs rely on direct prediction → shallow reasoning.
- RL-based methods (Med-R1, MedVLM-R1) bias toward reward tokens but lack reasoning emergence.
- Manual chain-of-thought (CoT) annotation is costly and inconsistent.
- Goal: Automatically generate reliable multimodal CoT data through collaborative reasoning search.

Contributions

- MICS: A collaborative multi-model search that builds reliable step-by-step reasoning data.
- MMRP: A large multimodal medical dataset covering 12 imaging modalities and 20 body systems.
- Chiron-o1: A curriculum-trained model with strong in-domain and out-of-domain reasoning.
- Results: Achieves state-of-the-art performance across multiple medical benchmarks.

MMRP Dataset Construction

- Part 1: Text QA (60K+ clinical cases) — basic domain knowledge.
- Part 2: Image–text alignment — real clinical imaging findings.
- Part 3: MICS-generated multimodal CoTs — stepwise reasoning supervision.
- Coverage: 12 imaging modalities × 20 body systems.
- Note: Enables curriculum learning from simple → complex reasoning.



MICS: Concept & Intuition

- Core idea: simulate mentor guiding interns; verify reasoning through collaboration.
- Mentor models (GPT-4o, Gemini 2.5, Qwen2.5-VL-72B) → propose reasoning paths.
- Intern models (Qwen2-VL-7B, InternVL3-8B) → complete reasoning based on mentor hints.
- Judge (DeepSeek-V3) → compare intern answers with ground truth.

“If most interns succeed, mentor’s reasoning path is valid — retained for dataset.”

The principle of MICS

The search proceeds iteratively as mentors and interns collaborate:

① Collaborative Search: mentors propose next reasoning step.

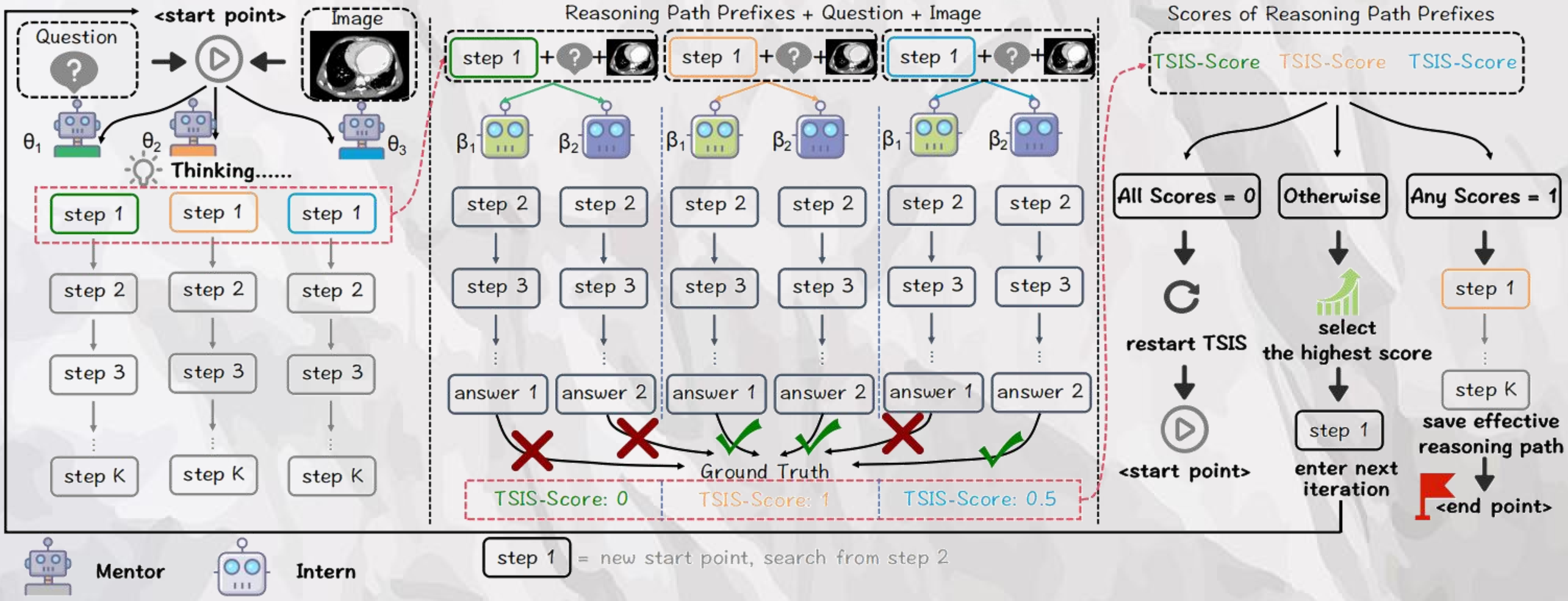
② Intern Evaluation: each intern completes reasoning → produces answer ã.

MICS-Score = (# interns correct) / (# interns total).

③ Path Selection: keep highest-score prefix for next iteration.

④ Early Stopping: restart if all scores=0, terminate if full-score achieved.

The principle of MICS



"Multi-agent self-evaluation ensures coherent, non-hallucinated reasoning paths."

Stage-wise Curriculum

- Stage 1 – Text QA → build core medical knowledge.
- Stage 2 – Image-Text Alignment → visual grounding.
- Stage 3 – MICS CoT Data → emergent reasoning.

- Base: InternVL3-8B | LoRA fine-tuning | AdamW lr=4e-5
  72 GPU-hours total (8 × A100) | mixed precision (bfloat16).

Experimental Results

Table 1: **Main Results on Medical VQA Benchmarks.** Our model achieves SOTA performance across various benchmarks. **Bold** denotes the highest score, and underline denotes the second-highest.

Methods	VQA-RAD	SLAKE	PathVQA	PMC-VQA	MMMU(H&M)	AVG
Close-Source SOTA						
Gemini-1.5-Pro [49]	60.3	72.6	70.3	52.3	47.9	60.7
Gemini-2.5-Pro [12]	71.3	80.5	73.9	**61.1**	**57.1**	68.8
GPT-4o-mini [23]	55.8	50.4	48.7	39.6	–	48.6
GTP-4o [23]	54.2	50.1	59.2	40.8	–	52.1
Open-Source SOTA						
LLaVA-v1.5-7B [33]	54.2	59.4	54.1	36.4	38.2	48.5
LLaVA-v1.6-13B [33]	55.8	58.9	51.9	36.6	39.3	48.5
LLaVA-v1.6-34B [33]	58.6	67.3	59.1	44.4	48.8	55.6
Yi-VL-34B [73]	53.0	58.9	47.3	39.5	41.5	48.1
Qwen-VL-Chat [2]	47.0	56.0	55.1	36.6	32.7	45.5
Medical MLLM						
LLaVA-Med [28]	51.4	48.6	56.8	24.7	36.9	43.7
Med-Flamingo [38]	45.4	43.5	54.7	23.3	28.3	39.1
RadFM [66]	50.6	34.6	38.7	25.9	27.0	35.4
GMAI-VL [30]	66.3	72.9	–	54.3	51.3	61.2
HuatuoGPT-Vision-7B [7]	63.8	74.5	59.9	52.7	49.1	60.0
HuatuoGPT-Vision-34B [7]	68.1	76.9	63.5	58.2	54.4	64.2
Reasoning Medical Model						
Med-R1	55.9	55.1	53.3	45.8	32.7	48.6
MedVLM-R1	61.4	65.9	55.2	44.8	35.5	52.6
ChestX-Reasoner	70.9	70.0	66.7	38.5	49.5	59.1
InternVL3-2B	68.3	65.9	65.2	49.1	38.4	57.4
Chiron-o1-2B	75.4	**85.3**	70.3	54.3	42.1	65.5$^{+8.1}$
InternVL3-8B	73.1	71.1	67.9	53.2	52.1	63.5
Chiron-o1-8B	**76.8**	83.2	**74.0**	57.5	54.6	**69.2**$^{+5.7}$

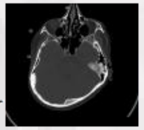
Table 2: **Results on Medical Reasoning Benchmarks.** ACC represents accuracy, and MICS-Score refers to the evaluation metric described in Equation 4. $*$ indicates pure text reasoning models.

Model	MMRP (Pure Text)	MedXpertQA_MM (Reasoning)	MedXpertQA_MM (Understanding)	MMRP (Reasoning)		
	ACC	ACC	ACC	ACC	Bert-Score	MICS-Score
MedReason*	79.2	–	–	–	–	–
HuatuoGPT-o1*	85.1	–	–	–	–	–
Med-R1	72.7	20.1	20.8	28.1	83.4	22.5
MedVLM-R1	77.5	21.7	20.0	31.2	83.5	23.5
Chiron-o1-2B	90.6	19.8	23.1	43.8	88.2	32.2
Chiron-o1-8B	**92.1**	**23.3**	**25.1**	**58.4**	**90.4**	**49.4**

Qualitative Results

user

Based on the imaging findings of abnormal flattening of the left parieto-occipital bones in this 20-year-old male patient, what is the most likely diagnosis for this incidental finding? Please base your response on the keyframes of the relevant medical imaging modality (Modality 1: <image>) and reason step-by-step to answer the above question.



Ground Truth: Primary congenital plagiocephaly

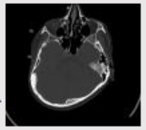
Med-R1

Step 1: The image is a CT scan of the head, which is a medical imaging modality used to visualize the internal structures of the body, including the brain and skull. The CT scan shows abnormal flattening of the left parieto-occipital bones, which are the bones that make up the back of the head and the back of the brain.
Step 2: The abnormal flattening of the left parieto-occipital bones suggests a potential pathology, such as a fracture, a tumor, or a congenital anomaly. However, without further clinical information or a detailed examination, it is difficult to determine the exact diagnosis.
Step 3: The Chief complaint is an incidental finding, which means that the patient has no specific symptoms or signs of a specific condition. This could indicate that the abnormal flattening of the left parieto-occipital bones is a result of a normal anatomical variation or a harmless incidental finding.
Step 4: The most likely diagnosis for this incidental finding is a normal anatomical variation or a harmless incidental finding. This is because the patient has no specific symptoms or signs of a specific condition, and the abnormal flattening of the left parieto-occipital bones is not associated with any specific symptoms or signs.

The final answer is: A normal anatomical variation or a harmless incidental finding. ✗ **(Shallow Thinking)**

user

Based on the imaging findings of abnormal flattening of the left parieto-occipital bones in this 20-year-old male patient, what is the most likely diagnosis for this incidental finding? Please base your response on the keyframes of the relevant medical imaging modality (Modality 1: <image>) and reason step-by-step to answer the above question.



Ground Truth: Primary congenital plagiocephaly

Chiron-o1

Step 1: Identify the key imaging finding and its location. The image shows an abnormal flattening of the left parieto-occipital bones. This observation is the primary focus of the case since it leads to the incidental finding.
Step 2: Consider the patient's age and gender. A 20-year-old male is within the typical age range for conditions that might cause such cranial abnormalities. Flat skull deformities, particularly in younger individuals, can be associated with specific syndromes or developmental variations.
Step 3: Analyze the potential causes of parieto-occipital bone flattening. In a young adult, this type of cranial deformation could be due to several reasons, including head trauma, prolonged external pressure on the skull, or genetic conditions affecting skull development. However, given the absence of any history of trauma or external pressure provided in the case information, these factors are less likely.
Step 4: Evaluate the possibility of plagiocephaly. Plagiocephaly is a condition characterized by an asymmetrical (usually flattened) head shape, often affecting one side of the parietal or occipital region. It can be congenital (primary plagiocephaly) or acquired (secondary plagiocephaly). Given the imaging findings of abnormal flattening of the left parieto-occipital bones, and considering the patient's age where secondary causes are less probable, primary congenital plagiocephaly becomes a strong candidate for diagnosis.
Step 5: Rule out other differential diagnoses. While other conditions like craniosynostosis (premature fusion of skull sutures) can lead to abnormal head shapes, the specific finding of parieto-occipital flattening without evidence of suture fusion suggests that plagiocephaly is more consistent with the observed imaging features.

The final answer is: Primary congenital plagiocephaly. ✓ **(Deep and Reasonable Thinking)**

Conclusion & Takeaways

- MICS: First collaborative framework for generating reliable medical Chain-of-Thought data.
- MMRP: Comprehensive multimodal reasoning dataset enabling curriculum-based training.
- Chiron-o1: Interpretable and generalizable medical MLLM achieving state-of-the-art results.

Check out our paper and poster!

https://arxiv.org/abs/2506.16962

Poster: Fri 5 Dec 11 a.m. PST — 2 p.m. PST & Exhibit Hall C,D,E (San Diego)

Questions: manglu3935@126.com