Learning to Be A Doctor: Searching for Effective Medical Agent Architectures

Yangyang Zhuang, Wenjia Jiang, Jiayu Zhang, Ze Yang, Joey Tianyi Zhou, Chi Zhang

AGI Lab, Westlake University, Henan University, Affiliated Hospital of Xuzhou Medical University, Xuzhou Medical University, Nanyang Technological University, IHPC, Agency for Science, Technology and Research, Singapore, CFAR, Agency for Science, Technology and Research, Singapore

August 29, 2025

Outline

Background: Large Language Models in Medicine

- Large Language Models (LLMs) are transforming various fields, including medicine.
- Their ability to handle interdisciplinary knowledge and complex tasks makes them ideal for medical applications.
- In healthcare, LLMs can enhance diagnostic accuracy, streamline workflows, and reduce the burden on professionals.

The Problem: Static Medical Workflows

- Current medical agent systems rely on static, manually crafted workflows.
- These workflows lack flexibility to adapt to diverse diagnostic requirements and emerging clinical scenarios.
- This rigidity limits scalability and effectiveness in real-world medical environments.

Our Core Contribution: Automated Framework Design

- Proposed the first fully automated framework for designing medical multi-agent systems using LLMs.
- Introduced hierarchical search space for dynamic workflow evolution.
- Developed self-improving architecture search algorithm guided by diagnostic feedback.

Methodology: Graph-Based Workflow Representation

- Medical workflows represented as graph-based structures with nodes and edges.
- Nodes categorized into basic nodes (LLM interaction) and tool nodes (external tools).
- Hierarchical search space enables three levels of modifications: node-level, structural-level, and framework-level.

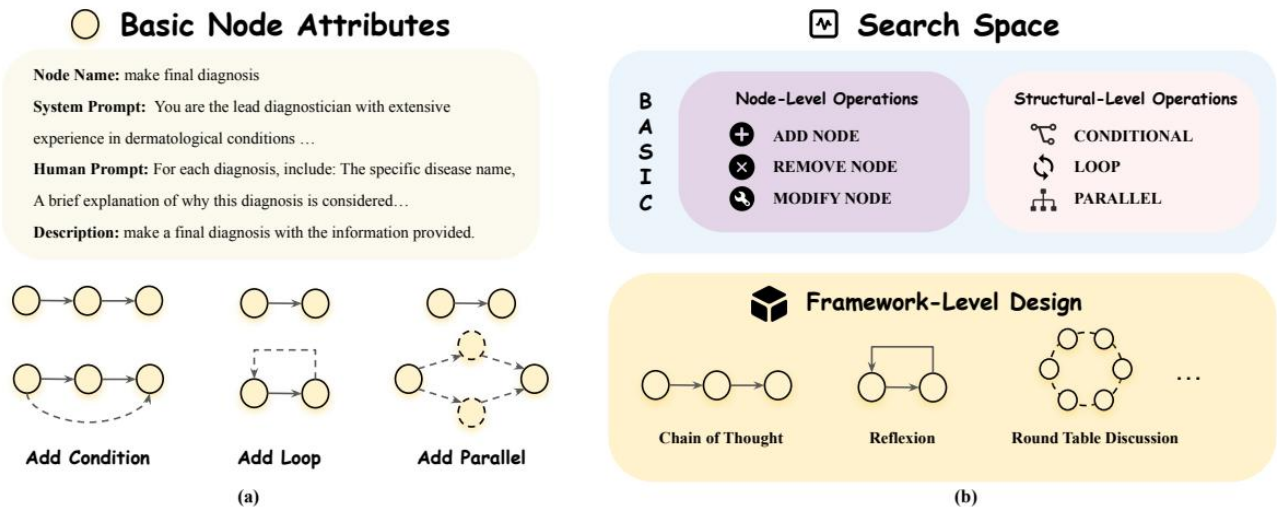

Figure: Node attributes and hierarchical search space for workflow design.

Workflow Evolution Process

- The framework iteratively refines workflows based on diagnostic feedback.
- LLM analyzes diagnostic errors, identifies root causes, and generates actionable suggestions.
- Suggestions are validated and implemented to optimize workflows over multiple iterations.

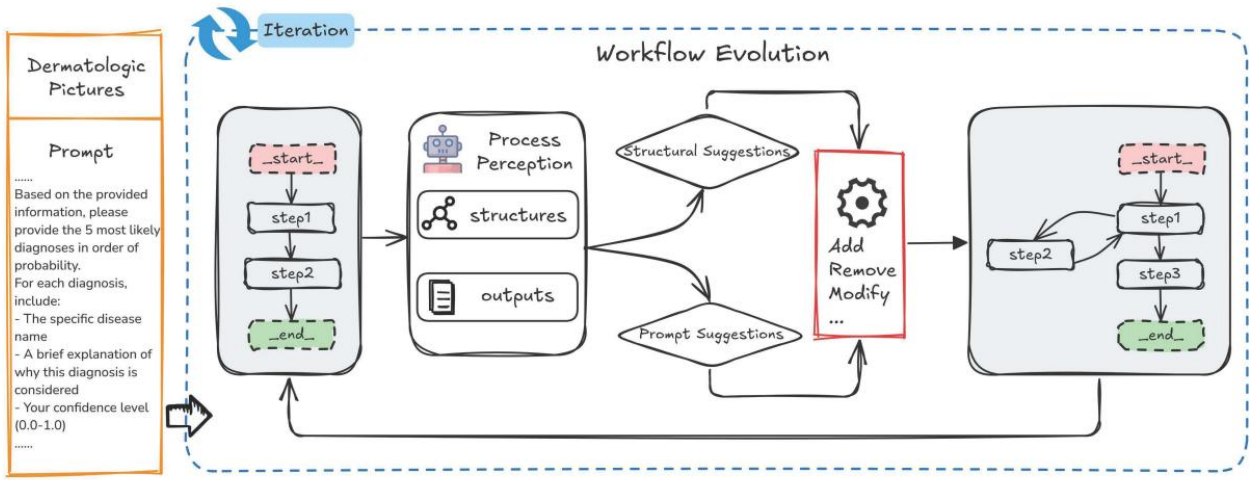

Figure: Iterative framework for optimizing workflows in medical diagnosis.

Experimental Setup: Datasets and Metrics

- Experiments conducted on two dermatological image classification datasets: Skin Concepts and Augmented Skin Conditions.
- Evaluation metrics include Top-k accuracy and consensus accuracy (cons@64).
- Baseline methods include Chain of Thought and Round Table frameworks.

Key Results: Diagnostic Accuracy

- Significant improvements across all evaluation metrics.
- Top-1 accuracy improved from 20.27% to 29.28% on Skin Concepts dataset.
- Achieved 90.83% Top-1 accuracy on Skin Conditions dataset.

Key Results: Diagnostic Accuracy

Table: Top-k diagnostic accuracy (%) of different methods using GPT-4o, GPT-4o-mini and Claude 3.5 Sonnet on Skin Concepts and Skin Conditions.

LLM	Method	Skin Concepts Accuracy (%)			Skin Conditions Accuracy (%)		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
GPT-4o	IO	20.27	30.63	36.04	50.83	78.33	86.67
	CoT (Wei et al., 2022)	18.47	28.83	33.78	55.83	76.67	82.50
	Round Table (Chen et al., 2024c)	21.17	27.93	32.43	45.83	75.83	80.83
	Ours	**29.28**	**40.09**	**50.45**	**90.83**	**95.00**	**100.00**
GPT-4o-mini	IO	11.71	20.72	23.87	27.50	69.17	80.83
	CoT (Wei et al., 2022)	6.31	15.32	24.32	22.50	65.00	84.17
	Round Table (Chen et al., 2024c)	10.81	19.82	23.42	25.83	70.00	78.33
	Ours	13.51	21.62	24.77	45.83	74.17	85.83
Claude 3.5 Sonnet	IO	17.12	24.77	26.13	40.00	68.33	75.83
	CoT (Wei et al., 2022)	14.86	22.07	24.32	36.67	63.33	73.33
	Round Table (Chen et al., 2024c)	15.77	22.52	26.58	35.00	66.67	74.17
	Ours	**28.83**	**35.14**	**38.29**	**95.83**	**98.33**	**99.17**

Ablation Study: Component Analysis

Table: Ablation results showing Top-k accuracies when specific operations are disabled.

Operation	Top-1 Acc. (%)	Top-3 Acc. (%)	Top-5 Acc. (%)
Add Tool Node	21.62 (-7.66)	30.18 (-9.91)	36.94 (-13.51)
Modify Node Prompt	19.37 (-9.91)	27.93 (-12.16)	33.78 (-16.67)
Remove Node	28.83 (-0.45)	41.44 (+1.35)	50.90 (-0.45)

Conclusion and Future Directions

- Introduced first automated framework for medical multi-agent system design.
- Achieved significant improvements in diagnostic accuracy and robustness.
- Future work includes broader medical domain adoption and integration with emerging technologies.

Questions & Discussion

- Thank you for your attention!
- Questions and feedback are welcome.