NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM
IJCAI 2025 GUANG ZHOU
The 2025 International Joint Conference on Artificial Intelligence

Tan Tock Seng HOSPITAL

ID: DM 87

A Smart Multimodal Healthcare Copilot with Powerful LLM Reasoning

Xuejiao Zhao[1,2*] Siyan Liu[1,2*] Su-Yin Yang[3] Chunyan Miao[1,2,†]

[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), NTU, Singapore

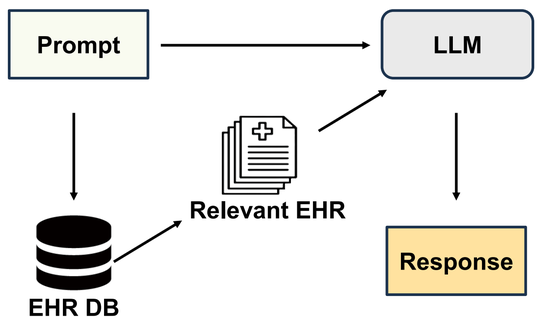
[3] Tan Tock Seng Hospital & Woodlands Health, Singapore

(*Co-first Author †Corresponding Author)

Retrieval-augmented Generation & Knowledge Graph

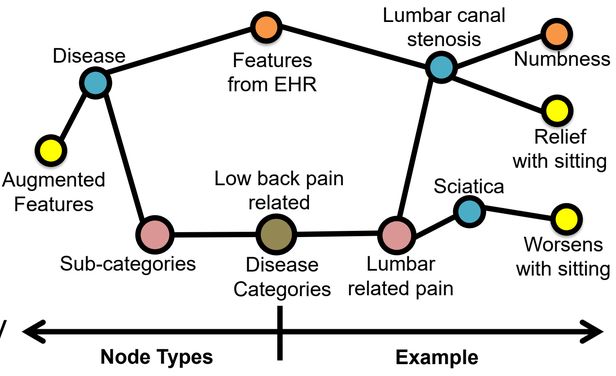
RAG (Retrieval-Augmented Generation):

- **Retrieves** relevant **local info** for **answer generation**
- **Improves** factuality and context-awareness of responses of LLMs
- **Suffers** from inaccuracies and vagueness due to heuristic-based approaches

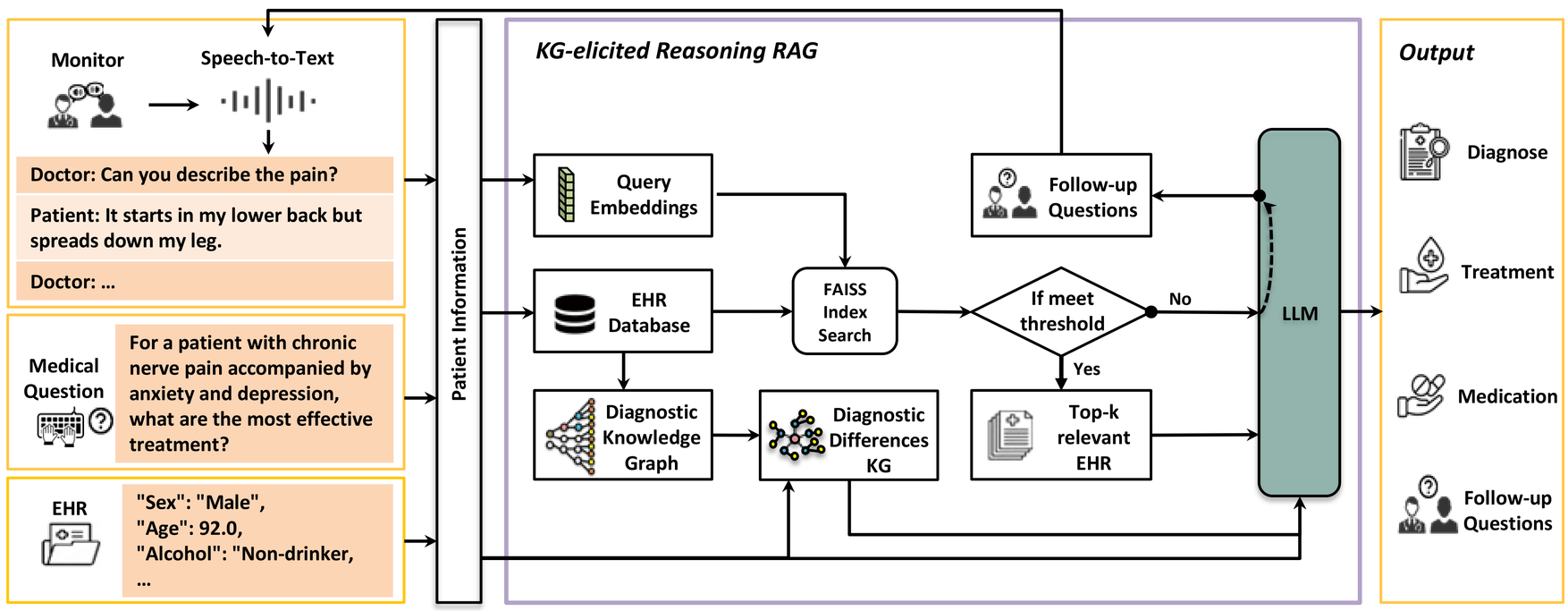


KG (Knowledge Graph):

- **Represents** medical entities (e.g., diseases, disease categories) and their relations
- **Elicits reasoning** of LLMs
- **Distinguishes** similar diseases and enhances diagnostic accuracy



Our Approach



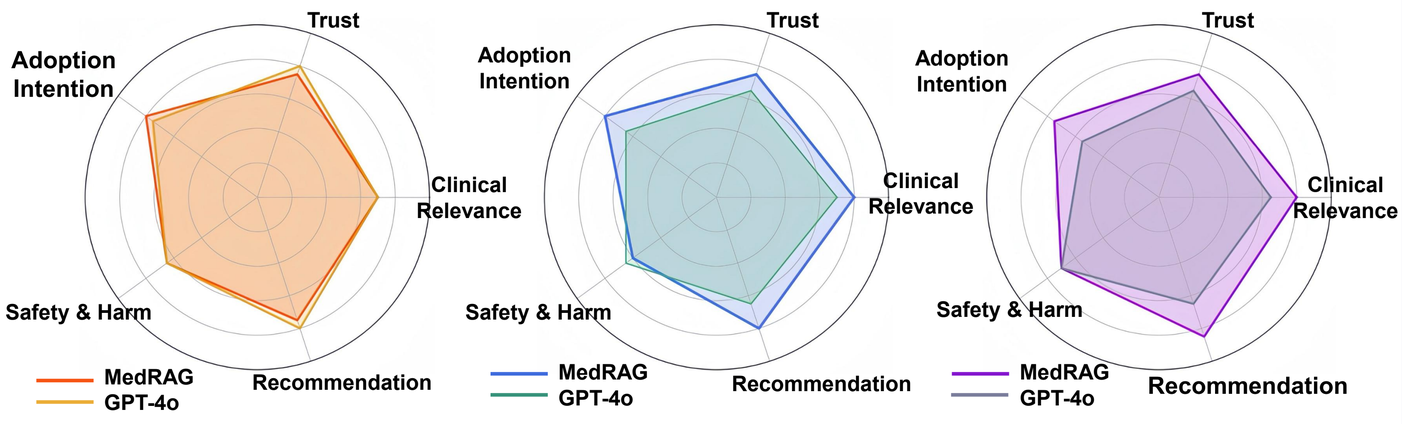
1. Diagnostic KG Construction: **A four-tier diagnostic KG** is constructed through disease clustering, hierarchical aggregation and LLM augmentation from EHR database.

2. Diagnostic Differences KG Searching: **Symptoms** are matched to the diagnostic KG via clinical feature decomposition, matching, and upward traversal to identify key **diagnostic differences**.

3. KG-elicited LLM reasoning: Based on the patient information, diagnostic difference KG and retrieved EHRs to reason precise **diagnostic suggestions and proactive questioning.**

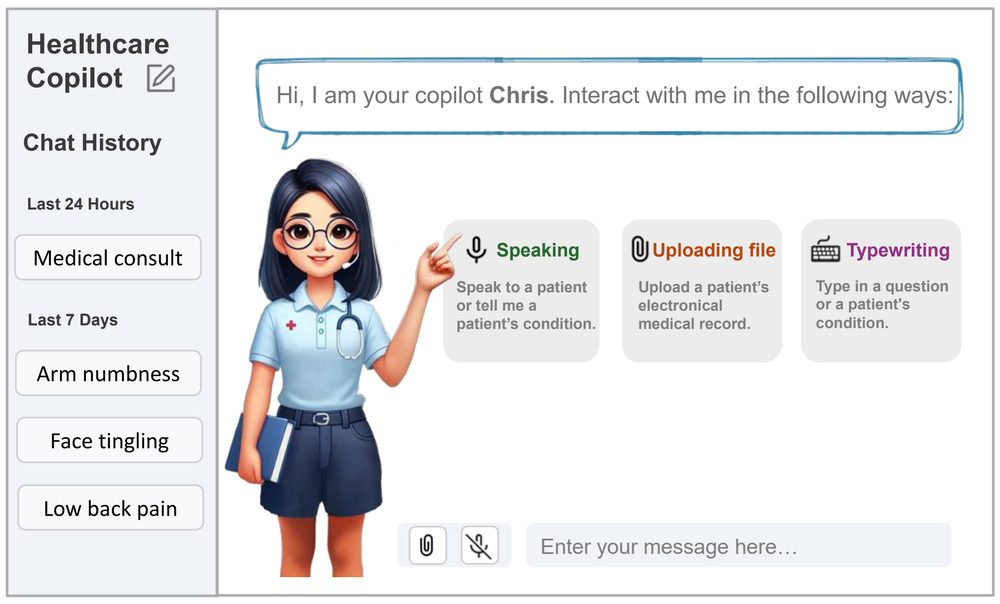
Evaluations and UI

Method	Model	CPDD			DDXPlus		
		L1	*L2*	*L3*	*L1*	*L2*	*L3*
Baselines	Naive RAG + COT	75.47	54.72	43.40	79.28	71.89	56.84
	FS-RAG	64.71	49.02	45.10	78.18	68.20	51.40
	FLARE	54.84	48.39	45.16	71.09	56.70	31.02
	FL-RAG	65.45	50.91	49.09	**90.12**	83.32	66.78
	DRAGIN	78.72	59.57	40.42	80.51	70.83	50.24
	SR-RAG	73.58	60.38	54.72	78.65	70.28	52.16
Ours	MedRAG	**79.25**	**75.47**	**66.04**	88.65	**83.46**	**68.01**



◆ **Outperform SOTA RAG methods on objective results**

◆ **Result of doctor evaluation (Human Factor Criteria)**



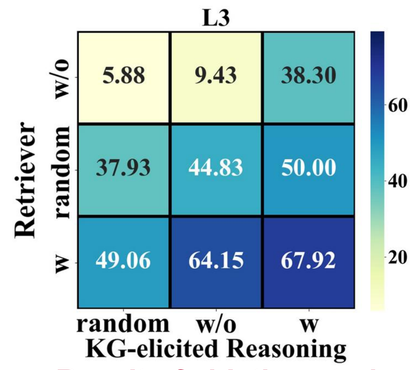
◆ **UI of MedRAG – Healthcare Copilot**

Backbone LLM	Modal	L1	L2	L3
GPT-4o	text	91.87	81.78	73.23
GPT-4o	voice	88.23	78.43	70.58
GPT-3.5-turbo	text	70.56	68.68	50.57
GPT-3.5-turbo	voice	64.70	60.78	45.09

◆ **Evaluation of different modal on CPDD**

Manifestation Masking Ratio	*L1*	*L2*	*L3*
100%	60.38	56.60	52.83
66.6%	69.39	67.35	55.10
33.3%	71.43	67.35	61.22
0%	79.25	75.47	66.04

◆ **Proactive diagnostic questioning**



◆ **Result of ablation study**

Demo on Bilibil

Code of MedRAG