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

THE WEB CONFERENCE ACM

The 2025 ACM Web Conference

Tan Tock Seng HOSPITAL

MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot

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

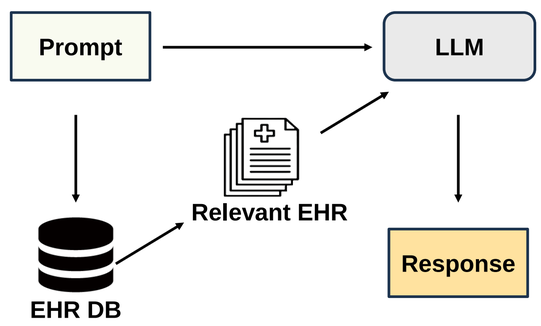
[3] Tan Tock Seng Hospital & Woodlands Health, Singapore

(* Co-first Author)

Retrieval-augmented Generation & Knowledge Graph

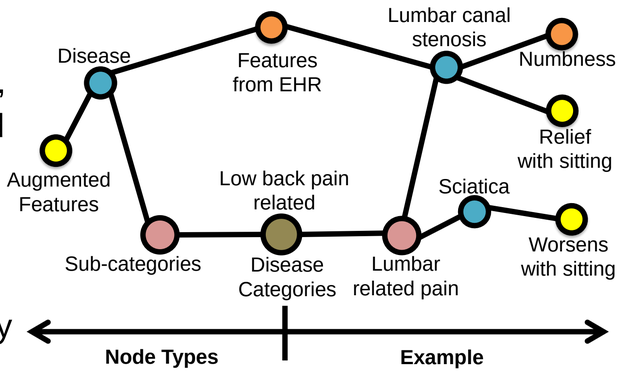
RAG (Retrieval-Augmented Generation):

- **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

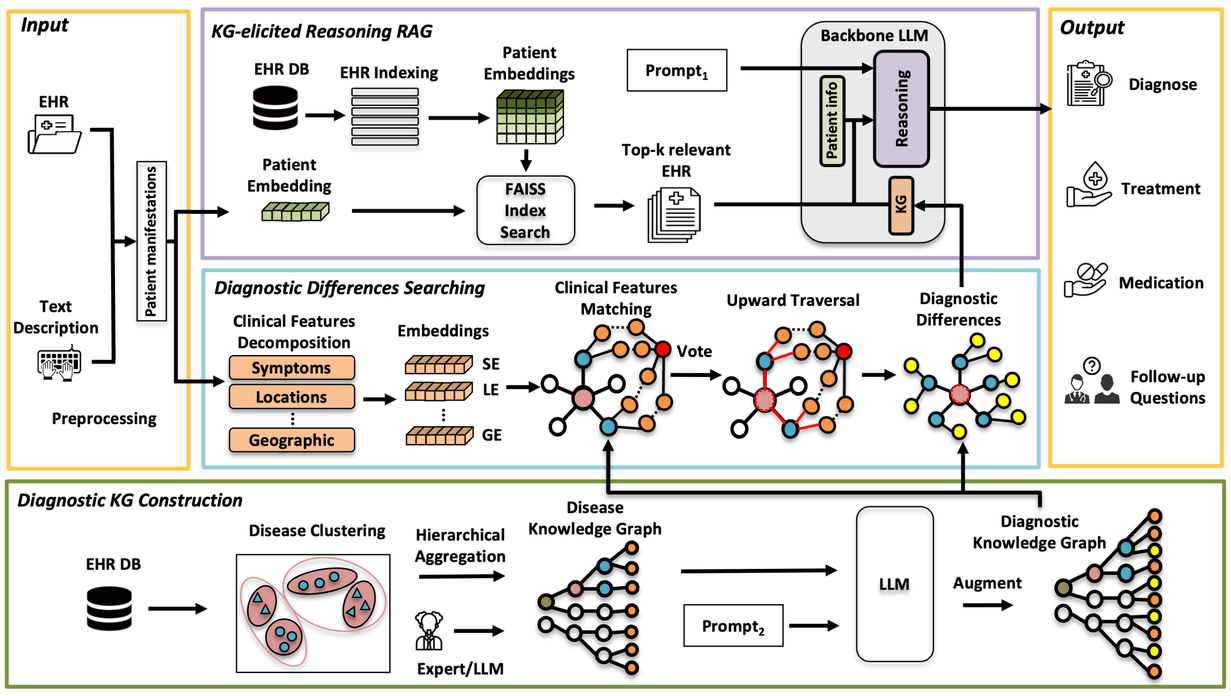


KG (Knowledge Graph):

- **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 Knowledge Graph (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 :

- **Patient symptoms** are matched to the diagnostic KG via clinical feature decomposition, matching, and upward traversal to identify key **diagnostic differences**.

3. KG-elicited Reasoning RAG :

- **Elicit LLM reasoning** based on the diagnostic difference KG, retrieved similar EHRs, and patient information to generate precise **diagnostic suggestions and proactive diagnostic questioning.**

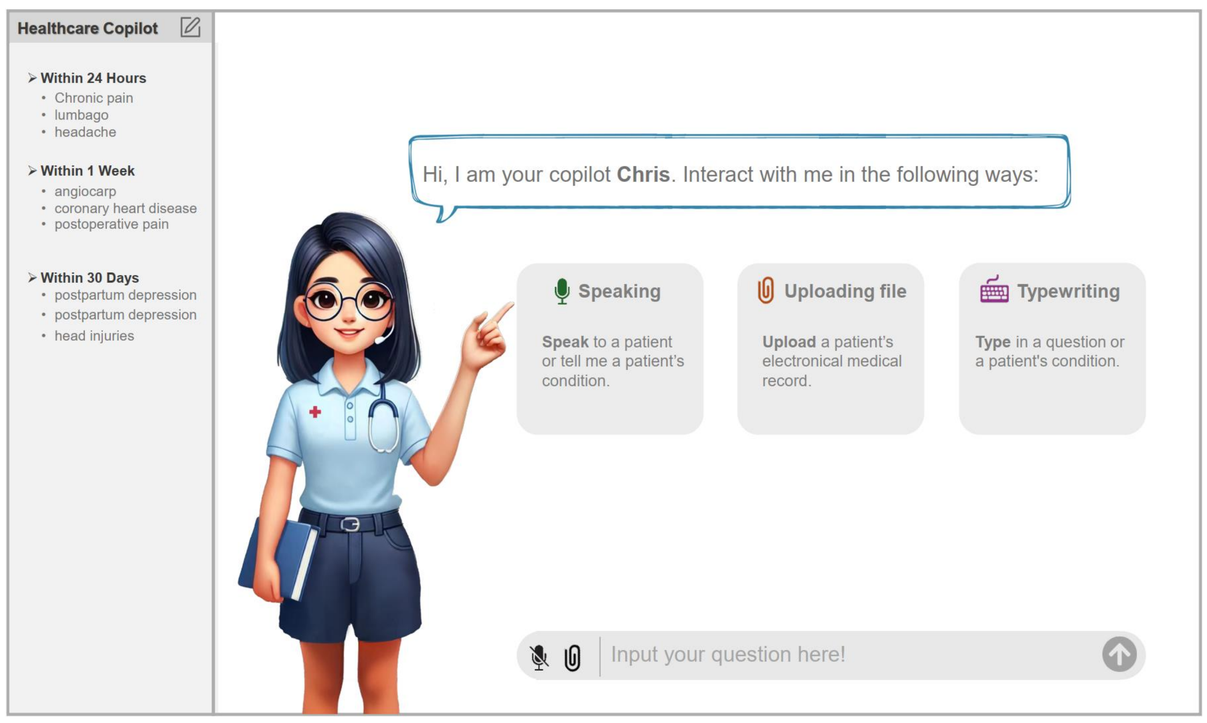
Evaluations and UI

	Backbone LLMs	Size	w/o KG-elicited Reasoning			w/ KG-elicited Reasoning		
			L1	*L2*	*L3*	*L1*	*L2*	*L3*
Open-source Models	Mixtral-8x7B	13B	60.38	32.08	22.34	84.62	82.69	63.46
	Qwen-2.5	72B	66.04	41.51	39.62	80.36	73.21	64.29
	Llama-3.1-Instruct	8B	75.47	54.72	43.40	79.25	75.47	66.04
	Llama-3.1-Instruct	70B	86.79	67.92	56.60	86.79	**83.02**	71.70
Closed-source Models	GPT-3.5-turbo	-	83.02	56.60	45.28	70.56	68.68	50.57
	GPT-4o-mini	-	88.68	67.92	56.60	85.85	75.00	60.38
	GPT-4o	-	**90.57**	**71.70**	**60.38**	**91.87**	81.78	**73.23**

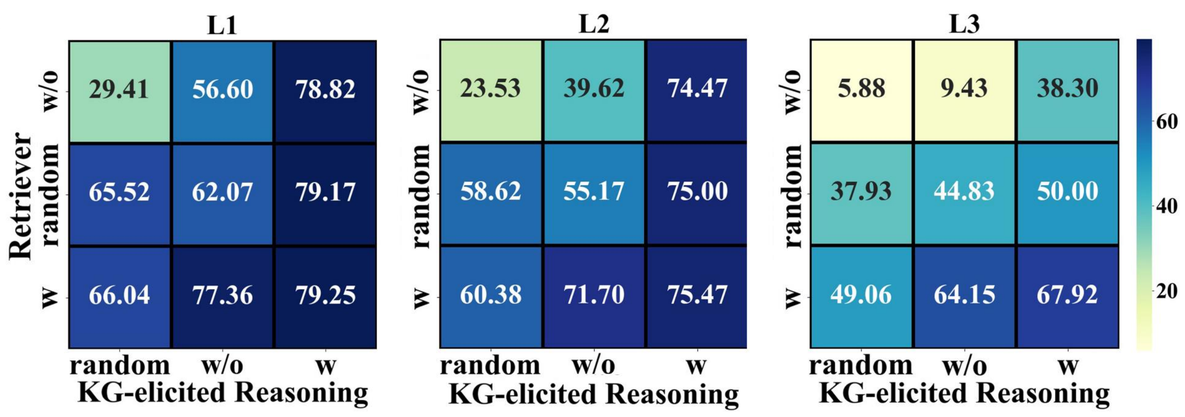
◆ **Performance of MedRAG on different LLM w and w/o KG-elicited reasoning**

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**



◆ **UI of MedRAG – Healthcare Copilot**



◆ **Ablation study of KG-elicited reasoning (Llama-3.1-Instruct 8B &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**

Demo of MedRAG