MedJudgeRAG: Option-Wise Evidence Judgment with Dynamic Knowledge Graphs for Medical MCQA

Seongwon Seo[1], Seung Hwan Cho[1], Young–Min Kim[1 2]

[1]Department of Industrial Data Engineering, Hanyang University, Seoul, South Korea

[2]School of Interdisciplinary Industrial Studies, Hanyang University, Seoul, South Korea

Code & Data

GFM Workshop@ ICML International Conference On Machine Learning

AML Applied Machine Learning Lab

Motivation

Vanilla RAG can perform worse than no retrieval in medical MCQA.

- RAG can improve domain-specific QA by supplying external evidence. However, Medical MCQA requires more than simple fact recall.
- Vanilla RAG indiscriminately prepends retrieved documents to the prompt.
- When the retrieved context is noisy or scattered, vanilla RAG can even hurt performance compared with parametric inference.

Missing mechanisms

- Connecting evidence scattered across multiple retrieved documents.
- Judging whether the evidence supports, contradicts, or is insufficient for each option.
- Deciding how retrieved evidence should influence final answer selection.

Training

Distill structured reasoning. Let the KG guide, not dominate.

- A teacher LM generates structured reasoning traces.
  ① KG → ② option-wise evidence judgment → ③ decision → ④ answer
- A student LM is trained with supervised fine-tuning on teacher-generated traces.
- We use a segment-weighted cross-entropy loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\frac{1}{Z}\left(\lambda_g \sum_{t\in\mathcal{T}_g} \log p_\theta(y_t \mid y_{<t}, x) + \sum_{t\in\mathcal{T}_r} \log p_\theta(y_t \mid y_{<t}, x)\right)\right]$$

  - $Z = \lambda_g|\mathcal{T}_g| + |\mathcal{T}_r|$
  - $\mathcal{T}_g$, $\mathcal{T}_r$: KG/reasoning tokens; $|\mathcal{T}_g|$, $|\mathcal{T}_r|$: token counts; λ_g: KG loss weight.

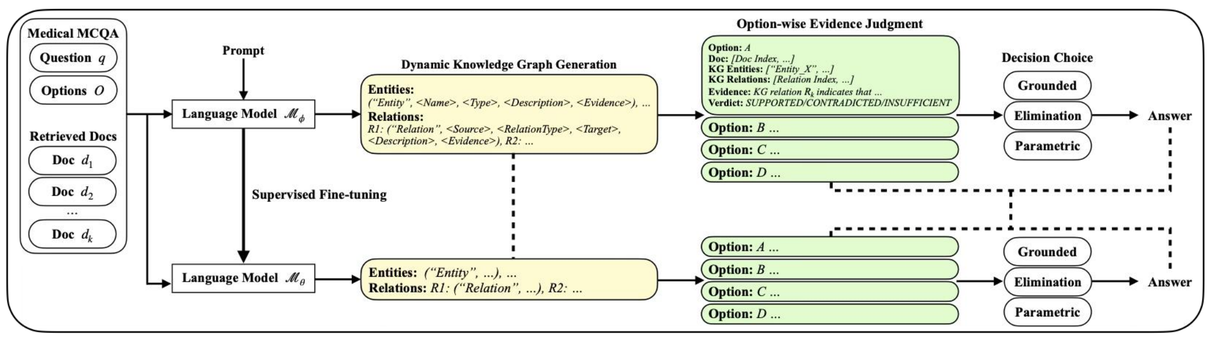
- KG-token losses receive weight λ_g, while reasoning tokens receive weight 1.0.
- The weighted-average normalization stabilizes loss scale across examples with different sequence lengths.

- Teacher LM: GPT-5.1 (gold answer not provided in the prompt)
- Student LMs: Mistral-7B-Instruct-v0.3, Meta-Llama-3-8B-Instruct
- Training
  - Setting: QLoRA + ZeRO-2, 3 epochs, max_seq_len = 8,192
  - Data: Mistral: 3,479 samples (3,148 train / 331 validation)
    Llama: 3,559 samples (3,222 train / 337 validation)

MedJudgeRAG Framework

Structure evidence. Judge options. Decide knowledge use.

- **MedJudgeRAG** turns retrieved documents into structured, option-wise evidence judgments before selecting an answer.



- **Step 1. Dynamic Knowledge Graph (KG) Construction**
  Retrieved documents are structured into a question-focused dynamic KG. Entities and relations are traced to source document indices.
- **Step 2. Option-wise Evidence Judgment**
  For each answer option, the model assigns one of three verdicts.
  (1) **SUPPORTED**: evidence supports the option as correct.
  (2) **CONTRADICTED**: evidence rules out the option.
  (3) **INSUFFICIENT**: evidence is insufficient to decide.
- **Step 3. Three-Way Knowledge Utilization Decision**
  Based on the verdict combination, the model selects how to use retrieved evidence.
  (1) **Grounded**: use SUPPORTED evidence.
  (2) **Elimination**: remove CONTRADICTED options.
  (3) **Parametric**: use model knowledge when all verdicts are INSUFFICIENT.

Main Results

Same retrieval, better evidence use.

Backbone	Method	MedQA	MedMCQA	Avg
Mistral	Parametric	52.55	46.26	49.41
	Vanilla RAG	50.59	41.57	46.08
	Ours (Explicit)	59.54	50.37	54.96
	Ours (Implicit)	60.88	50.16	55.52
Llama	Parametric	60.64	54.86	57.75
	Vanilla RAG	49.88	45.61	47.75
	Ours (Explicit)	63.94	55.39	59.67
	Ours (Implicit)	66.93	57.33	62.13

- Vanilla RAG can underperform parametric inference, indicating that retrieval alone is insufficient.
- The model must determine how to use retrieved evidence for each answer option.
- Across both Mistral and Llama backbones, **MedJudgeRAG** improves over vanilla RAG using the same retriever and retrieved documents.

Ablation

KG works best as supervision, not output.

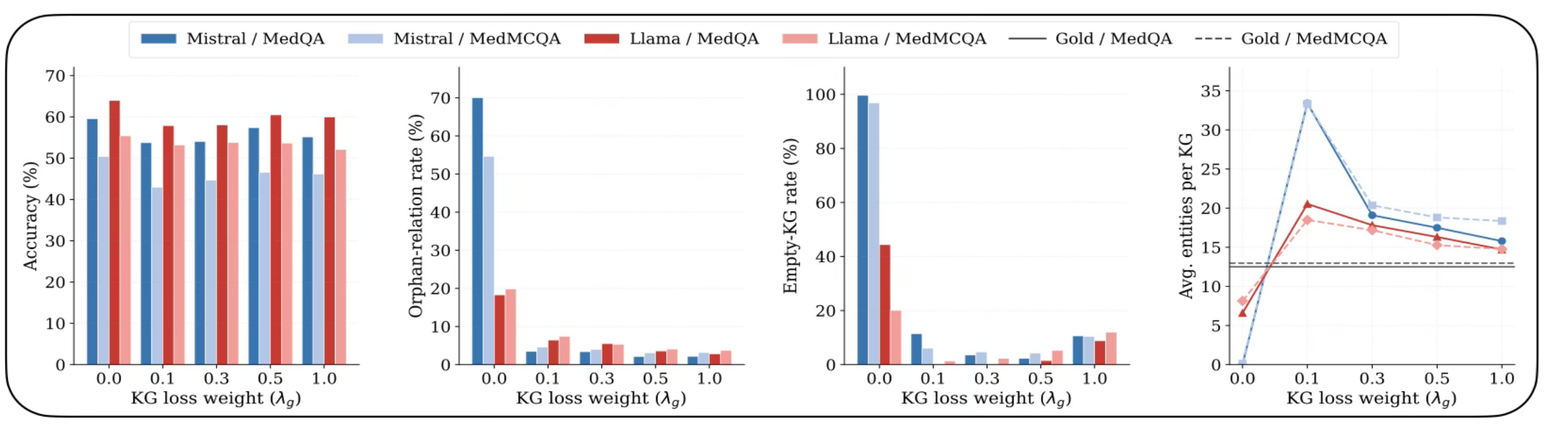
Graph-conditioned supervision improves answer accuracy, while mandatory explicit KG decoding can reduce generation robustness.

Bench.	Method	KG loss weight (λ_g) 0.0	0.1	0.3	0.5	1.0	w/o KG
Mistral							
MedQA	Exp.	**59.54**	53.73	54.05	57.34	55.15	55.15
	Imp.	60.88	**61.19**	60.17	59.23	60.49	
MedMCQA	Exp.	**50.37**	42.96	44.68	46.52	46.12	47.57
	Imp.	**50.16**	49.39	48.96	49.06	48.82	
Llama							
MedQA	Exp.	**63.94**	57.89	58.05	60.49	59.94	62.29
	Imp.	**66.93**	65.59	64.65	63.32	63.16	
MedMCQA	Exp.	**55.39**	53.14	53.77	53.65	52.09	55.87
	Imp.	**57.33**	56.13	56.01	55.77	54.82	

Backbone	Bench.	Exclusive correct Exp.	Imp.
Mistral	MedQA	143.6	200.0
	MedMCQA	479.4	611.0
Llama	MedQA	125.2	184.6
	MedMCQA	427.8	528.4

- Explicit KG decoding is not always optimal. The dynamic KG is most effective as graph-conditioned supervision during training, rather than as a mandatory explicit output at inference time.
- Implicit KG decoding skips KG generation at inference but retains the reasoning behavior learned from KG-conditioned traces. It wins in 19 / 20 settings.
- Complementary, Not Inferior

  Explicit and Implicit decoding answer different questions correctly. Explicit recovers cases Implicit misses; Implicit recovers more cases Explicit misses. On average, Implicit wins by avoiding the KG generation bottleneck. Still, Explicit is not strictly inferior.



- Increasing λ_g generally improves KG structural consistency but does not necessarily improve answer accuracy.
- Explicit decoding can over-generate long KGs, consume output budget, and fail to transition to the <ANALYSIS> reasoning phase. This leads to more invalid outputs that do not select A/B/C/D.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25492127).