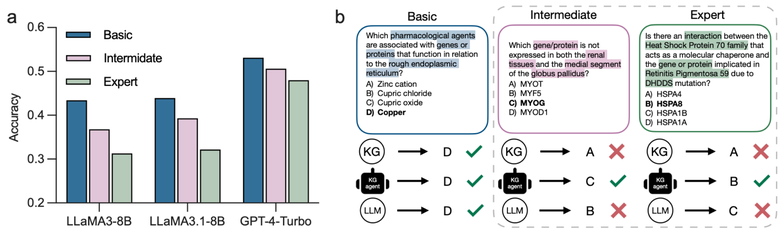
KGARevion: An AI Agent For Knowledge-Intensive Biomedical QA

Xiaorui Su[1], Yibo Wang[2], Shanghua Gao[1], Xiaolong Liu[2], Valentina Giunchiglia[3], Djork-Arné Clevert[4], Marinka Zitnik[1]

[1]Harvard University, [2]University of Illinois Chicago, [3]Imperial College London, [4]Pfizer

xiaorui_su@hms.harvard.edu marinka@hms.harvard.edu

LLMs don't reason effectively in complex scenarios



- LLMs trained in general-purpose data struggle to solve medical problems that require specialized knowledge in the domain.
- LLMs often do not reason effectively in complex scenarios, where successful inference requires recognizing and reasoning about dependencies between multiple interrelated medical concepts within a single question and interpreting highly similar yet semantically distinct biomedical entities with precision.

KGARevion is robust multiple choice selectors

Question: A 29-year-old woman presents to the clinic with a 6-month history of progressive weakness and muscle pain. She has experienced difficulty walking and has had several falls in the past month. Her symptoms have progressed despite taking ibuprofen and acetaminophen. Physical examination reveals muscle atrophy in her upper and lower extremities. Laboratory tests show elevated creatine kinase levels and a positive test for Human Immunodeficiency Virus (HIV). What is the most likely diagnosis?

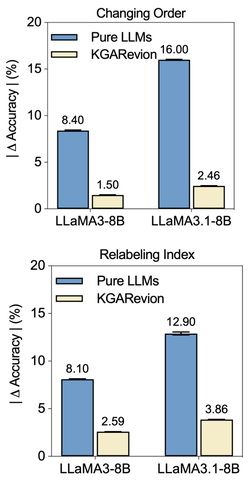
Answer Options:
A: Myopathy **B:** Polymyositis **C:** Dermatomyositis **D:** Neuromuscular junction disorder

Changed order of answer options

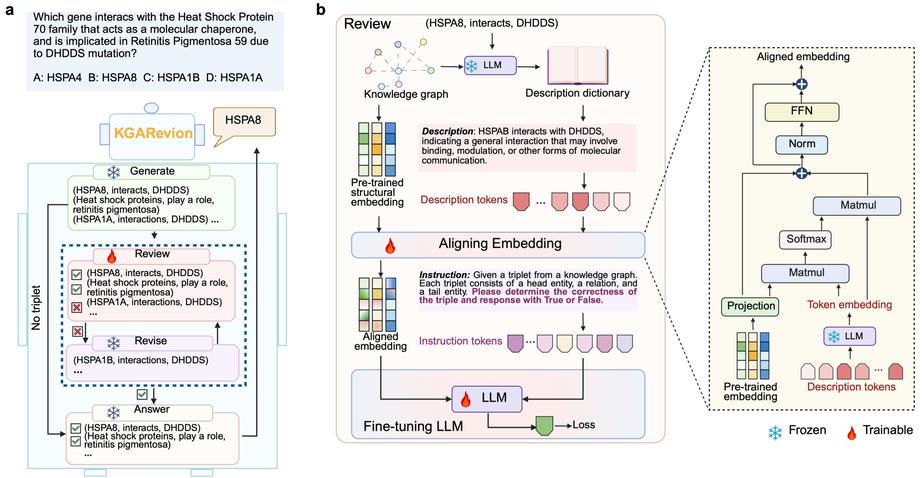
Answer Options:
- **C:** Dermatomyositis
- **A:** Myopathy
- **D:** Neuromuscular junction disorder
- **B:** Polymyositis

Relabeled index

Answer Options:
- **G:** Myopathy
- **H:** Polymyositis
- **I:** Dermatomyositis
- **J:** Neuromuscular junction disorder



Overview of KGARevion agent

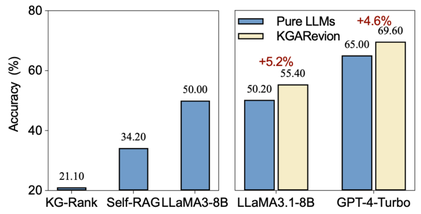


KGARevion, composed by four actions, integrates both LLM-generated knowledge and structured KG-based validation, improving accuracy and robustness in intensive medical QA

- *Generate*: generate triplet related to the input question
- *Review*: assess the correctness of each generated triplet by grounding knowledge in KGs
- *Revise*: correct any triplet identified as incorrect
- *Answer*: produces the final answer based on the triplets verified by the *Review* action

Benchmark KGARevion on a newly dataset: AfriMed-QA

Inference directly on all expert-level multiple-choice questions in AfriMed-QA

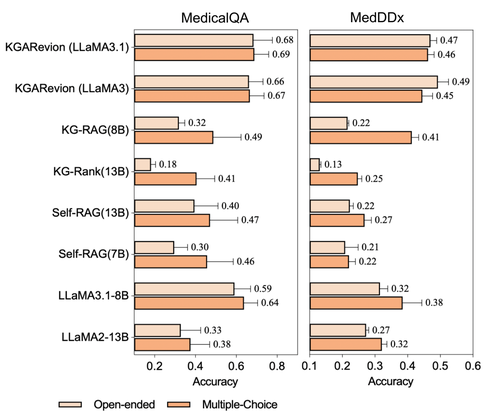


- Afrimed-QA dataset is sourced from over 500 clinical and non-clinical contributors across 16 countries and covers 32 clinical specialties.
- The dataset is sourced from healthcare systems whose data are not online, which means that none of the questions and answers can be in knowledgebase of the LLM

✓ **No leakage** in backbone LLMs. ✓ **No leakage** in adopted KGs. ✓ **No leakage** in KGARevion.

KGARevion excels in intensive biomedical QA

KGARevion performs best under Multiple-Choice and Open-ended settings



Multiple-Choice

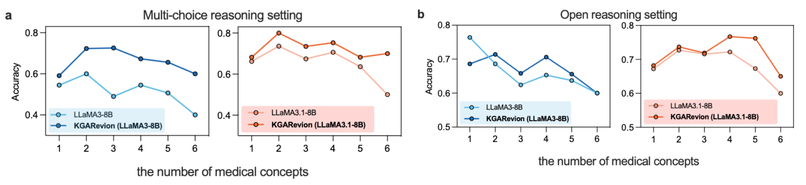
Which of the following best describes the structure that collects urine in body?

A. Bladder
B. Kidney
C. Ureter
D. Urethra

Open-ended

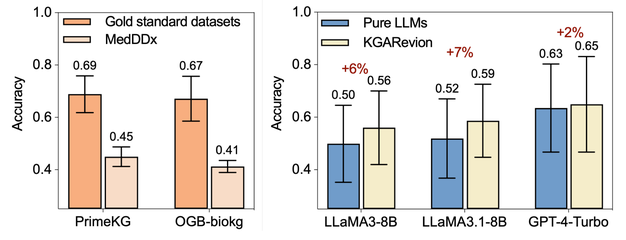
What best describes the structure that collects urine in body?

KGARevion improves the accuracy of LLMs with complex questions

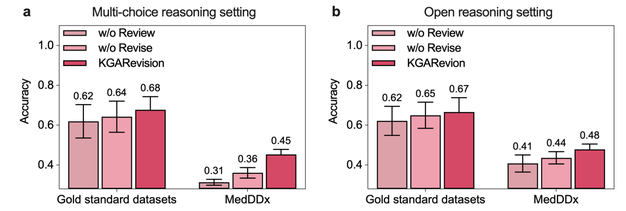


- More medical concepts contained in question, model complex the question is!

KGARevion is flexible with different KGs and LLMs



Does Review and Revise action matter?



Paper

Project Website

Code