Medical Question Answering for Critical Care Medicine

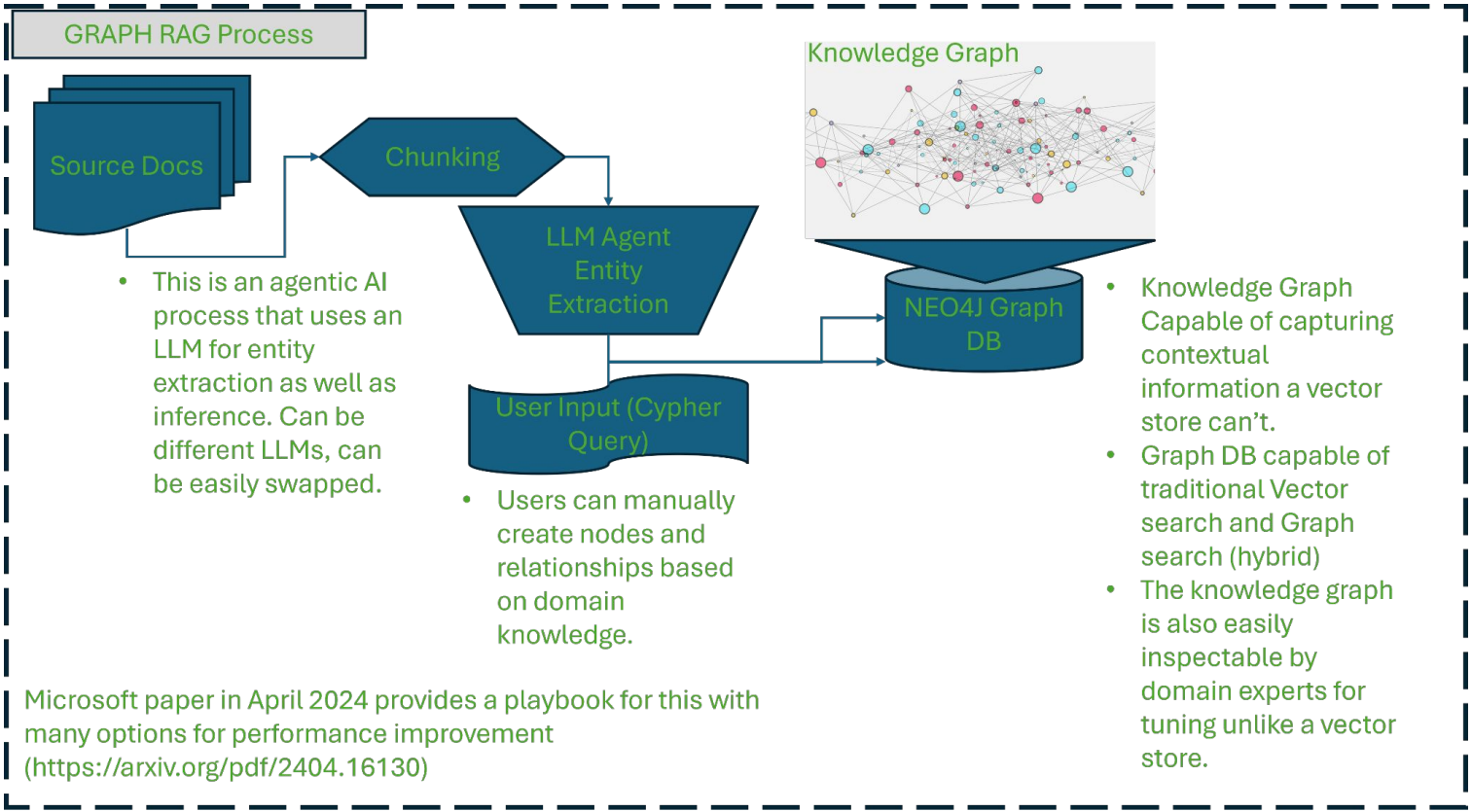
Progress and Milestones for Spring Semester

Jan. 16 2025

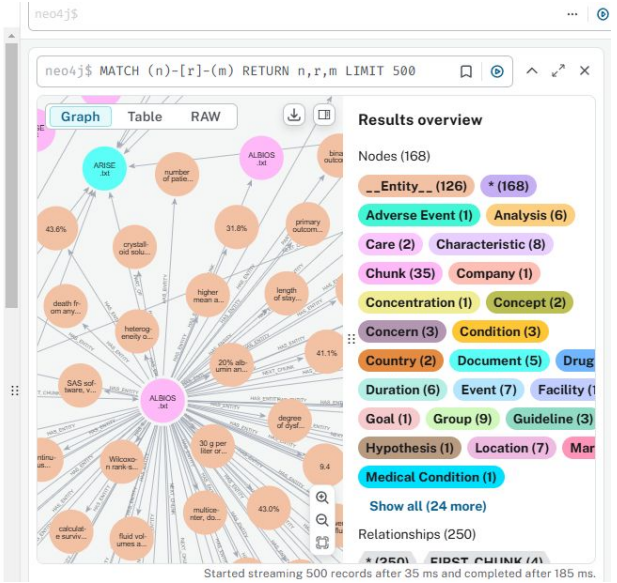
Outline

- Progress
  - Improving the GraphRAG
- Literature Review: Human Evaluation & Benchmarking Datasets
- Milestones for the semester

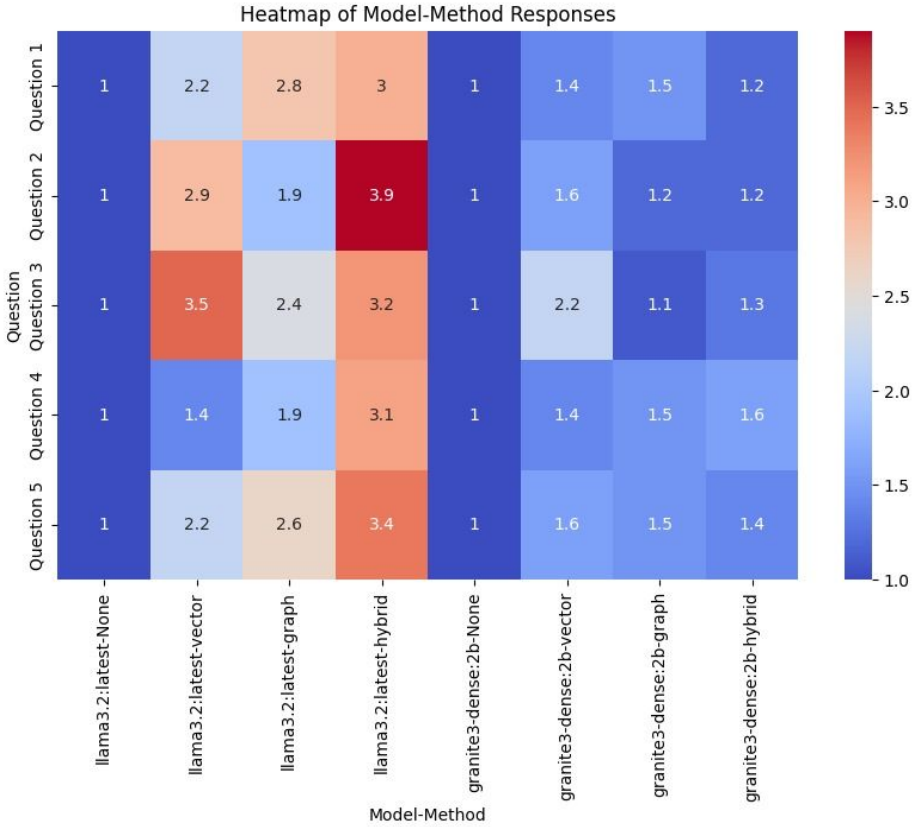
Quick review



Knowledge graph of ARDS and sepsis



Results across different models (show notebook)



Answer Evaluation

- Retrieval and Generation
- Instruction Following
- Structure and Coherence
- Comprehensiveness
- Diversity
- Empowerment
- Directness

Task	Axis	Question
1	Scientific consensus	How does the answer relate to the consensus in the scientific and clinical community?
2	Extent of possible harm	What is the extent of possible harm?
3	Likelihood of possible harm	What is the likelihood of possible harm?
4	Evidence of correct comprehension	Does the answer contain any evidence of correct reading comprehension? (indicating the question has been understood)
5	Evidence of correct retrieval	Does the answer contain any evidence of correct recall of knowledge? (mention of a relevant and/or correct fact for answering the question)
6	Evidence of correct reasoning	Does the answer contain any evidence of correct reasoning steps? (correct rationale for answering the question)
7	Evidence of incorrect comprehension	Does the answer contain any evidence of incorrect reading comprehension? (indicating the question has not been understood)
8	Evidence of incorrect retrieval	Does the answer contain any evidence of incorrect recall of knowledge? (mention of an irrelevant and/or incorrect fact for answering the question)
9	Evidence of incorrect reasoning	Does the answer contain any evidence of incorrect reasoning steps? (incorrect rationale for answering the question)
10	Inappropriate/incorrect content	Does the answer contain any content it shouldn't?
11	Missing content	Does the answer omit any content it shouldn't?
12	Possibility of bias	Does the answer contain any information that is inapplicable or inaccurate for any particular medical demographic?

Benchmarking Dataset - Existing QA Dataset

Dataset	Description
PubMedQA	**1,000 expert-labelled question–answer pairs**. The task is to produce a **yes/no/maybe** answer given a question together **with a PubMed abstract as context**. Each yes/no/maybe answer is **followed by a summary text, which was derived from the "conclusion" of the source research paper.**
LiveQA	Curated as part of the Text Retrieval Challenge (TREC) 2017. The dataset consists of **medical questions** submitted to the National Library of Medicine (NLM). The dataset also consists of **manually collected reference answers from trusted sources** such as the National Institute of Health (NIH) website. **Size** (development set/test set): **634/104**.
MedicationQA	commonly asked **consumer questions about medications**. **Size** (development set/test set): **NA/674**.
HealthSearchQA	**3,173 commonly searched consumer questions** (Question only). The dataset was **curated using seed medical conditions and their associated symptoms**.

Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge.*Nature* 620, 172–180 (2023). Doi: 10.1038/s41586-023-06291-2

Ideas for Experiments

- Model
  - Proposed methods (e.g. Graph-based)
  - Benchmarks (e.g. Microsoft GraphRAG)

- Evaluation
  - Human Evaluation (Small subset)
  - Automatic Measures
  - Correlation Analysis

- Dataset
  - Questions relevant 10-15 medical conditions derived from Wiki Journal Club
  - Sample Documents: Sampling Methodologies and Sample Size
  - Benchmarking Datasets

Milestones

Date	Description
Jan. 30	**Literature Review** (Graph-based Medical QA "B", Benchmarking "Ma" / Curated Dataset "Mu", Answer Evaluation"K"), **Draft Methods**
Feb. 13	**Proposal** (Human Rating Methodology, Dataset Curation, Answer Evaluation), **Preliminary Results** (Benchmarking dataset sample) → Complete Cost Estimation
Feb. 27	**Dataset Overview, Results** (sample benchmarking/ curated dataset), Human Rating Analysis → Finalize Human rating methodologies
Spring break	
Mar. 20	**Sample results, interpretation, troubleshooting** → **Finetune proposed solution**, complete all experiments
Apr. 3	Rehearsal: Project Presentation (Major Results and Interpretation)
Apr. 4	MIDS Capstone Symposium

PubMedQA

Example question:

Double balloon enteroscopy (DBE): is it **efficacious and safe in a community setting**?

Answer: Yes.

Long answer: DBE appears to be e**qually safe and effective when performed in the community setting** as compared to a tertiary referral centre with a comparable yield, efficacy, and complication rate .

BACK

LiveQA

Example question:

Could **second hand smoke contribute to or cause early AMD**?

Long answer:

Smoking increases a person's chances of developing AMD by two to five fold. Because the retina has a high rate of oxygen consumption, a **nything that affects oxygen delivery to the retina may affect vision. Smoking causes oxidative damage,** which may contribute to the development and progression of this disease. Learn more about why smoking damages the retina, and explore a number of steps you can take to protect your vision .

BACK

MedicationQA

Example question:

How does valium affect the brain?

Focus (drug): Valium.

Question type: Action.

Long answer:

Diazepam is a benzodiazepine that **exerts anxiolytic, sedative, muscle-relaxant, anticonvulsant and amnestic effects.** Most of these effects are **thought to result from a facilitation of the action of** gamma aminobutyric acid (**GABA**), **an inhibitory neurotransmitter** in the central nervous system.

BACK

HealthSearchQA

Example question:

How serious is atrial fibrillation?

Example question:

What kind of cough comes with Covid?

Example question:

Is blood in phlegm serious?

BACK