Medical Question Answering for Critical Care Medicine

Experimental Design and Semester Plan Revisited

Feb. 13 2025

Overview

- **Objective**
  - We aim to design and execute an **experiment to demonstrate our proposed methodology** for medical question answering for critical care medicine.
- **Outline**
  - Experiment design (Dataset, Medical Questions, Answer Annotation)
  - Output Evaluation (Relevance, Factuality, Synthesis)
  - Example Output
  - Weekly Milestones

Output Format

Example Question/ Claim:

Steroids improves survival and reversal of shock **in patients with septic shock**.

Supporting the claim:

Paper 1
Study Design and Methodology:
Study Population:
Interventions:
Comparator:
Outcomes:
Strengths and Weaknesses:
Key Findings and Conclusion:

Paper 2
…

Against the claim:

Paper 3
Study Design and Methodology:
Study Population:
…

Dataset

- 60 research papers related to 4 medical conditions (ARDS, Sepsis, Cardiac Arrest, Delirium) are collected for model build up and evaluation.
- All research papers included are associated with at least one medical condition **and** one medical question.

Medical Questions

4 Topics ARDS, Sepsis, Cardiac Arrest, Delirium

- Does **early administration** of **neuromuscular blocking agents** increases the ventilator free days?
- Patients with septic shock undergoing mechanical ventilation, did **continuous infusion of hydrocortisone** result in **lower 90-day mortality**?
- In patients with coma after out-of-hospital cardiac arrest, did **targeted hypothermia** lead to a **lower incidence of death by 6 months** than targeted normothermia?
- Was there a **difference between dexmedetomidine and midazolam in time at targeted sedation level** in mechanically ventilated ICU patients?

Ground Truth Annotation I

- Each research paper was **annotated** in terms of **relevance and stance against *all*** in medical questions (yes-no) or claim. The annotations can be used for **automatic output evaluation**

Relevance Annotation

	Q1	Q2	…
Paper 1	1 (relevant)	0 (irrelevant)	…
Paper 2	…	1 (relevant)	…
…	…	…	…

Stance Annotation

	Q1	Q2	…
Paper 1	1 (supporting)	0 (neutral)	…
Paper 2	…	-1 (against)	…
…	…	…	…

Ground Truth Annotation II

- For each research paper, we leveraged large language models (specifically, Claude from AWS Bedrock), to **generate summaries**. **After editing and proofreading by physicians** (KN, MA), the summaries will serve as **ground truth** for automatic/ human evaluation.

	Study Design and Methodology	Study Population	Interventions	…	Key Findings and Conclusion
Paper 1			…		
Paper 2			…		
…	…	…	…		

Evaluation

Retrieval

Compare retrieved docs and annotated docs

Factuality

Answer factually correct (PICO?)

Chunks factually correct

Synthesis

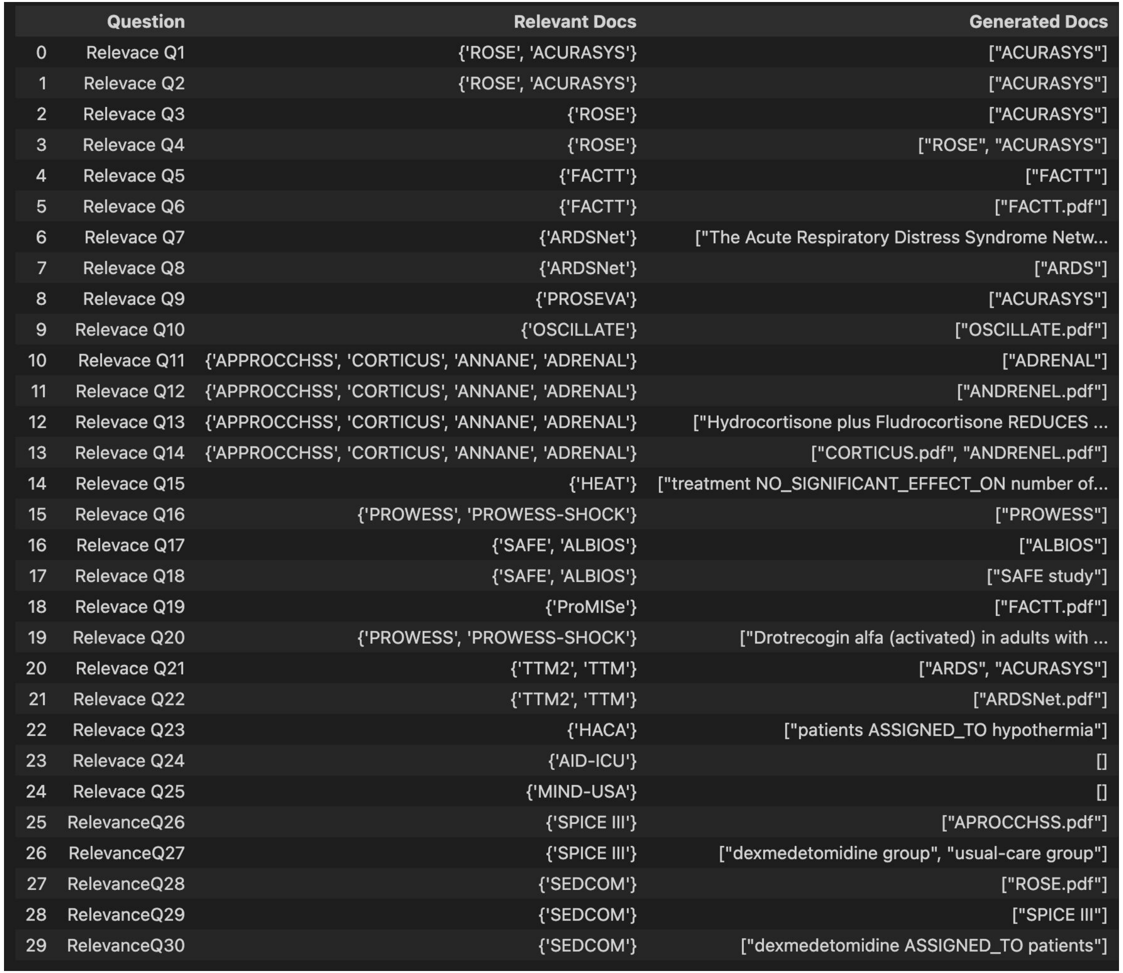
1. Granularity and extraction quality (PICO?)
2. Consistency (pass@1)
3. Scalability
4. Manuscript order
5. Contradictory information

Evaluation

Retrieval

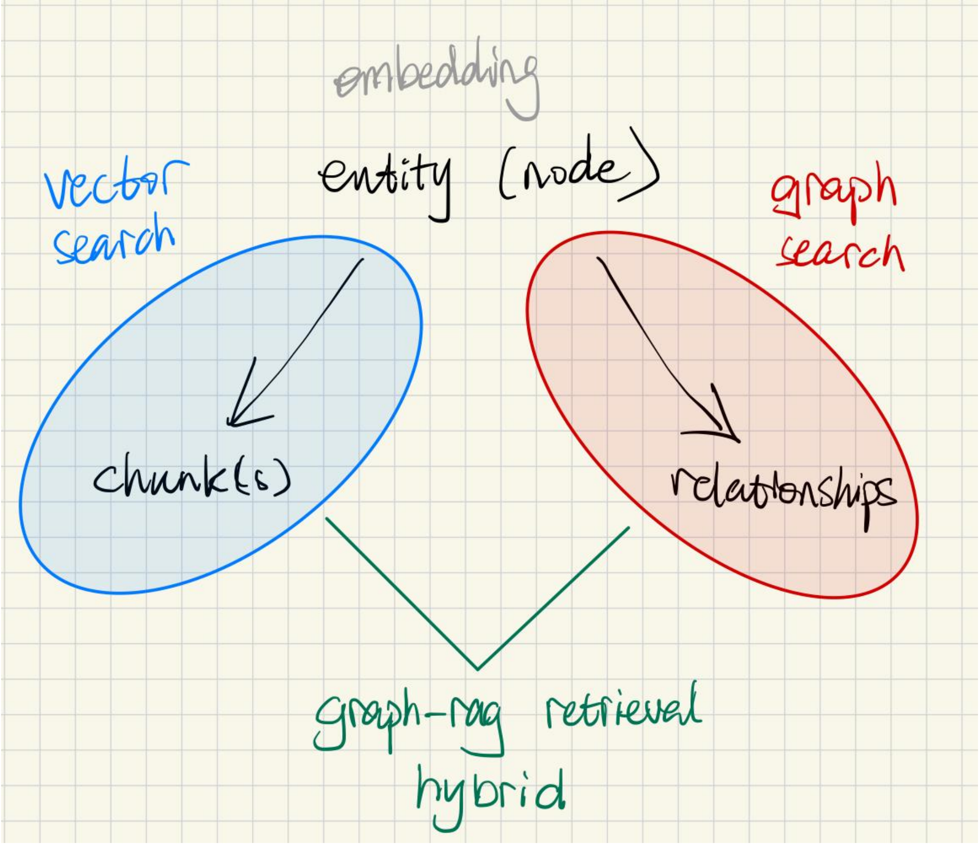
Compare retrieved docs and annotated docs

1. Demo
2. Potential issues
3. Github link



	Question	Relevant Docs	Generated Docs
0	Relevace Q1	{'ROSE', 'ACURASYS'}	["ACURASYS"]
1	Relevace Q2	{'ROSE', 'ACURASYS'}	["ACURASYS"]
2	Relevace Q3	{'ROSE'}	["ACURASYS"]
3	Relevace Q4	{'ROSE'}	["ROSE", "ACURASYS"]
4	Relevace Q5	{'FACTT'}	["FACTT"]
5	Relevace Q6	{'FACTT'}	["FACTT.pdf"]
6	Relevace Q7	{'ARDSNet'}	["The Acute Respiratory Distress Syndrome Netw...
7	Relevace Q8	{'ARDSNet'}	["ARDS"]
8	Relevace Q9	{'PROSEVA'}	["ACURASYS"]
9	Relevace Q10	{'OSCILLATE'}	["OSCILLATE.pdf"]
10	Relevace Q11	{'APPROCCHSS', 'CORTICUS', 'ANNANE', 'ADRENAL'}	["ADRENAL"]
11	Relevace Q12	{'APPROCCHSS', 'CORTICUS', 'ANNANE', 'ADRENAL'}	["ANDRENEL.pdf"]
12	Relevace Q13	{'APPROCCHSS', 'CORTICUS', 'ANNANE', 'ADRENAL'}	["Hydrocortisone plus Fludrocortisone REDUCES ...
13	Relevace Q14	{'APPROCCHSS', 'CORTICUS', 'ANNANE', 'ADRENAL'}	["CORTICUS.pdf", "ANDRENEL.pdf"]
14	Relevace Q15	{'HEAT'}	["treatment NO_SIGNIFICANT_EFFECT_ON number of...
15	Relevace Q16	{'PROWESS', 'PROWESS-SHOCK'}	["PROWESS"]
16	Relevace Q17	{'SAFE', 'ALBIOS'}	["ALBIOS"]
17	Relevace Q18	{'SAFE', 'ALBIOS'}	["SAFE study"]
18	Relevace Q19	{'ProMISe'}	["FACTT.pdf"]
19	Relevace Q20	{'PROWESS', 'PROWESS-SHOCK'}	["Drotrecogin alfa (activated) in adults with ...
20	Relevace Q21	{'TTM2', 'TTM'}	["ARDS", "ACURASYS"]
21	Relevace Q22	{'TTM2', 'TTM'}	["ARDSNet.pdf"]
22	Relevace Q23	{'HACA'}	["patients ASSIGNED_TO hypothermia"]
23	Relevace Q24	{'AID-ICU'}	[]
24	Relevace Q25	{'MIND-USA'}	[]
25	RelevanceQ26	{'SPICE III'}	["APROCCHSS.pdf"]
26	RelevanceQ27	{'SPICE III'}	["dexmedetomidine group", "usual-care group"]
27	RelevanceQ28	{'SEDCOM'}	["ROSE.pdf"]
28	RelevanceQ29	{'SEDCOM'}	["SPICE III"]
29	RelevanceQ30	{'SEDCOM'}	["dexmedetomidine ASSIGNED_TO patients"]

GraphRAG hybrid retrieval mechanism



Potential Issues

1. **Embedding dimensionality:**
   a. High -> sparsity
   b. Low -> lose semantic meaning
   c. Langchain 1564 vs Bob 384. Neo4j not sure
2. **Entity extraction:**
   a. NER (Named Entity Recognition)
   b. GPT model
   c. Langchain gpt-4 vs Bob gpt4. Neo4j not sure
3. **Cypher query for retrieval**
4. **Prompt design**

Milestones

- Will update weekly progress using google doc.

Date	Description
Feb. 14	Finalize experimental design (dataset, evaluation), Medical Questions Proposal.
Feb. 21	Answer and Medical questions done. Paper collection done with annotation. Relevance and factuality results done.
Feb. 28	Experimental results update (model refinement).
Mar. 7	Synthesis Experiment (paradigms in synthesis). Introduction write up
Mar. 14	Troubleshooting. Potentially adding documents. (Spring Break)
Mar. 21	Model refinement and troubleshooting. Result write up. Draft symposium presentation.
Mar. 28	Finalize symposium presentation. Discussion write up.

Backup

Outline

- Experimental Design
- Overview (M)
- Dataset Curation (M)
- Annotation (M)
- Medical Questions (K)
- Evaluation (Relevance, Factuality, Synthesis) (B)
- Preliminary Results
- Example Output (B)
- Weekly Milestone for the semester (S)