Developing A Medical Question Answering Model for Duke Department of Medicine

Bob Zhang, Keon Nartey, Suim Park, Yun-Chung (Murphy) Liu

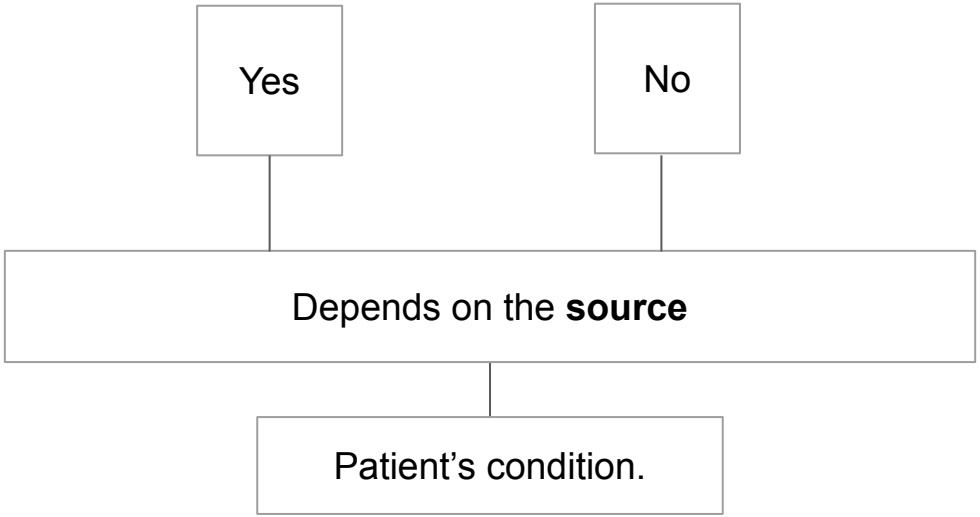
March 20 2025

Objective

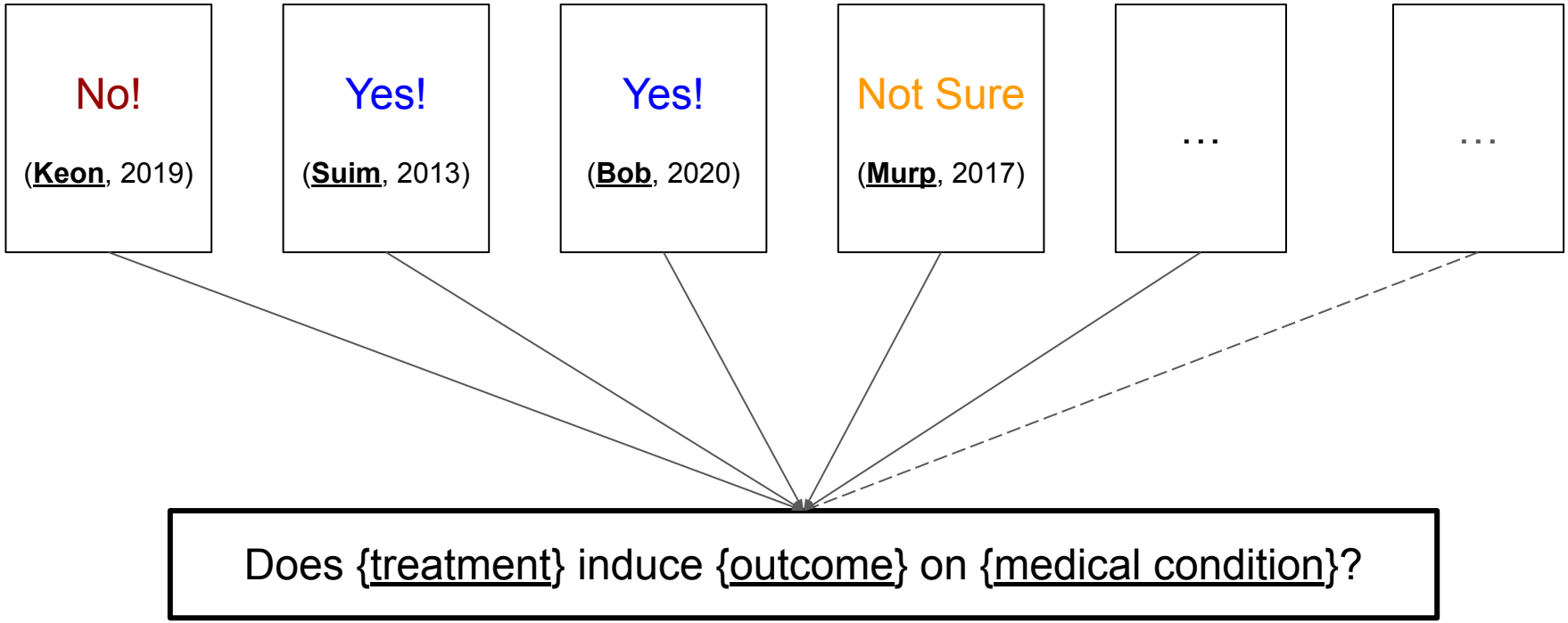
- **Background**
  - As the number of research papers and guidelines increase, it becomes a **tremendous burden for physicians** to stay up-to-date to **make informed decisions** faced with various clinical conditions.
- **Objective**
  - The project aims to develop of an **evidence-based medical question answering model** for critical care medicine

An Example Question

Do patients with severe **ARDS** (Acute Respiratory Distress Syndrome, a medical condition) being treated with **neuromuscular blocking** (medication) agents have increased **muscle weakness**? (Outcome)

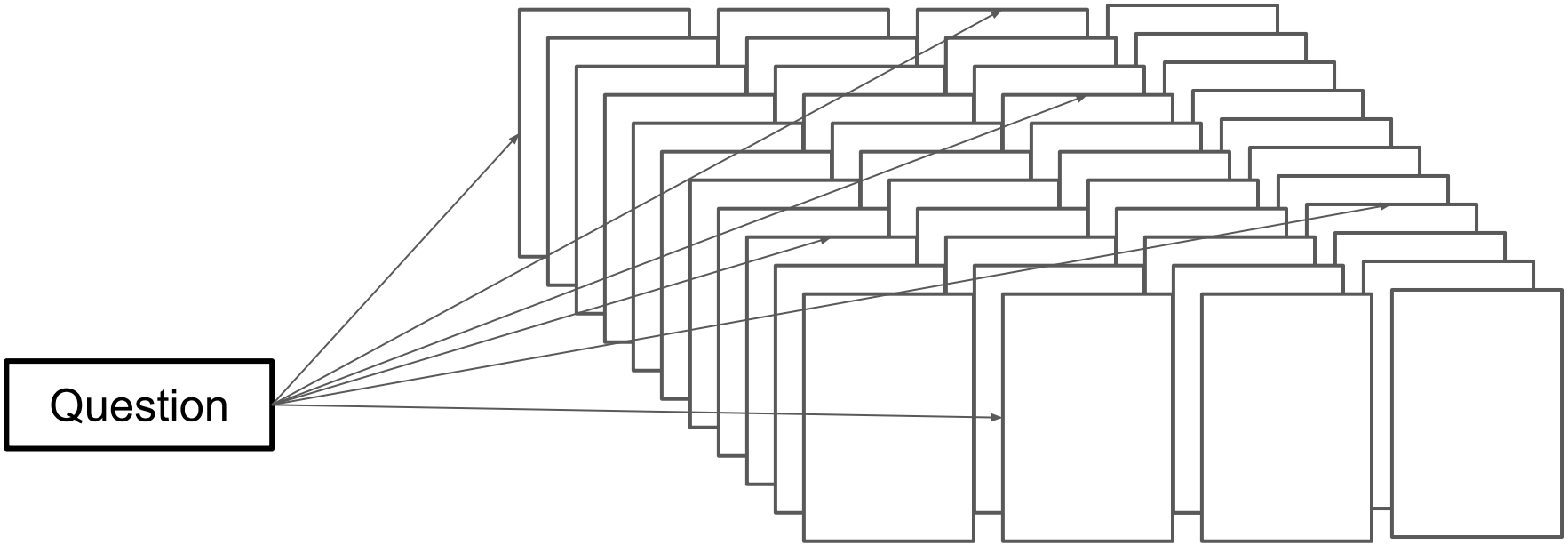


Medical Questions Answered by Research Papers



Quantity Counts

What happens when the amount of source document becomes HUGE?



Output Format

Medical Question/ Claim:

Do **steroids** **improves** **survival** and reversal of shock **in patients with septic shock**?

User Question

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:

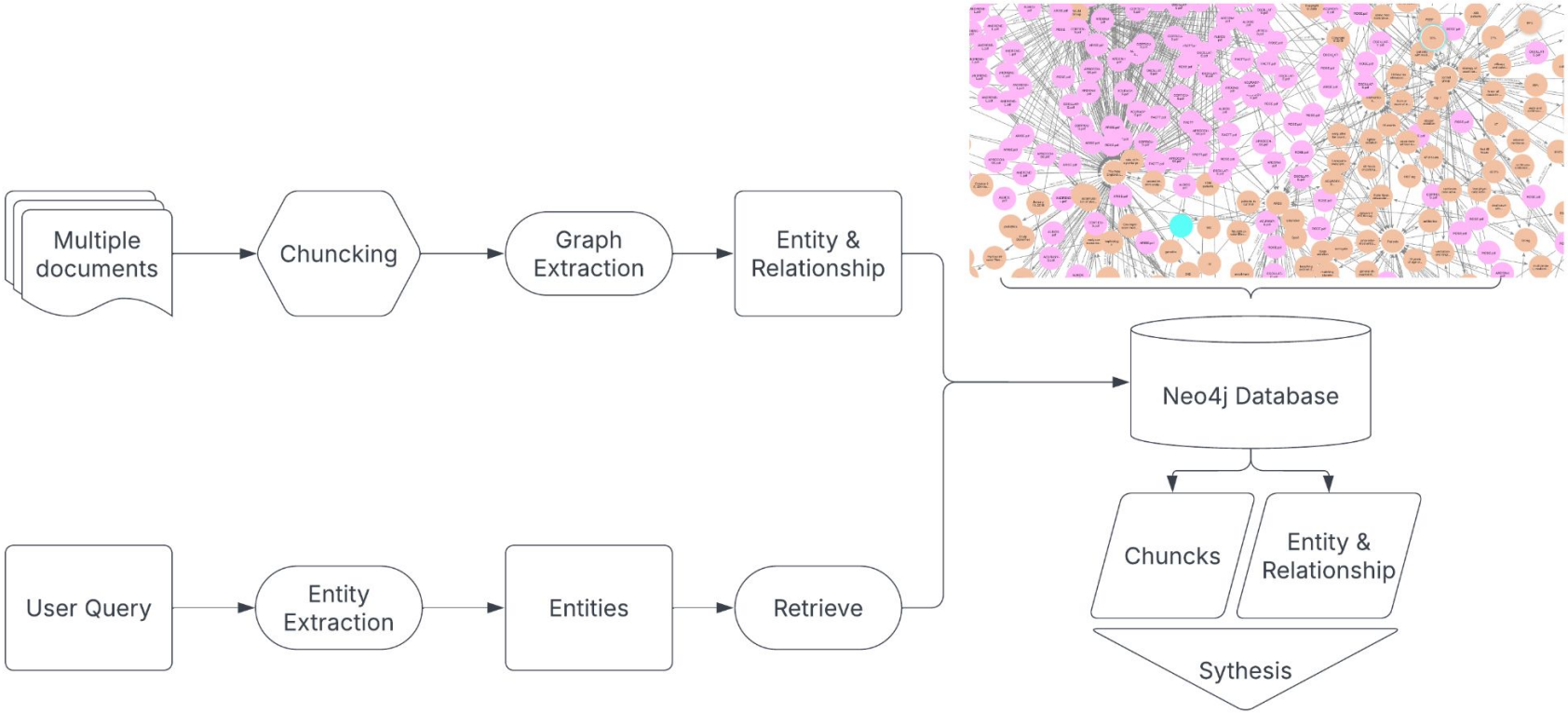
Paper 3
Study Design and Methodology:
Study Population:

Papers Supporting The Claim

Papers Against The Claim

A Summary of each paper

Retrieval Augmented Generation Plus Knowledge Graph



Dataset

- **Research Papers**
- 112 research papers related to any medical conditions are collected for model build up and evaluation. (34 papers from WikiJournal, 86 from PubMed, Google Scholar)
- **Medical Questions**
- 36 yes-no medical questions (12 for ARDS, 10 for sepsis, 6 for cardiac arrest, 7 for delirium, 1 for Sepsis *and* delirium)**.**

Ground Truth Annotation

- Each research paper was **annotated** in terms of **relevance and stance against *all*** in medical questions (yes-no) or claim.

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 model (Claude from AWS Bedrock), to **generate summaries**. **After editing and proofreading by physicians** (Keon), the summaries serve as **ground truth**.

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

Evaluation

- **Relevance**
  - Relevance of retrieved papers measured measured by accuracy, precision, recall, and F-1 score.
- **Factuality**
  - Summaries are evaluated on factual correctness and semantic similarities against ground truth using Ragas.

Results - Retrieval

	Recall	Precision	F-1	Accuracy
Naïve RAG (Suim)	.476	.560	.444	.896

* Mean value of all metrics from 36 questions

Issue - Hallucination

Example Question/ Claim:

← NON EXISTING PAPER

Does early administration of neuromuscular blocking agents increases the ventilator free days?

Supporting the claim:

Paper 2: **Impact of Early Neuromuscular Blockade on ARDS Mortality**

a. Authors: Moss et al.
b. Year: 2016
c. Journal: Critical Care Medicine
d. Research Question: Does early NMBA administration affect mortality in ARDS patients?
e. **Study Design and Methodology**: Cohort study analyzing data from a large database.
f. **Study Population**: 1,000+ ARDS patients.
g. **Interventions & Comparators**: Early NMBA administration vs. standard care.
h. **Outcomes**: Decreased mortality and increased ventilator-free days.
i. **Strengths and Weaknesses**: Large sample size is a strength; retrospective design is a weakness.
j. **Key Findings and Conclusion**: Supports the claim with significant findings on mortality and ventilator-free days.

Issue - Incorrect Stance

Equivocal or Inconclusive:
1. **Study 1**
   - **Authors:** Annane et al.
   - **Year:** Not specified
   - **Title:** "Effect of Treatment With Low Doses of Hydrocortisone and Fludrocortisone on Mortality in Patients With Septic Shock"
   - **Journal:** Not specified
   - **Research Question:** Examined the impact of low-dose corticosteroids on 28-day survival in patients with septic shock and adrenal insufficiency.
   - **Type of Study:** Placebo-controlled, randomized trial
   - **Methodology:**
     - Randomized patients to hydrocortisone and fludrocortisone vs. placebos.
     - Main outcome: 28-day survival in nonresponders to corticotropin.
   - **Key Findings:**
     - No significant difference in 28-day mortality between hydrocortisone and placebo.
     - No difference in mortality between responders and nonresponders to corticotropin.
   - **Conclusion:** Low-dose corticosteroids did not improve 28-day survival in septic shock patients with adrenal insufficiency.

Question: Does early application of high frequency oscillatory ventilation compared with ventilation strategy of low tidal volume decrease mortality?

Ground Truth: Supporting (1) / **Synthesis**: Equivocal or Inconclusive (0)

Synthesis Evaluation

	Study Design and Methodology
Ground Truth (ACURASYS)	- **Multicenter, randomized, double-blind**, placebo-controlled trial - March 2006 through March 2008 - 20 Intensive Care Units (ICUs) in France - Patients **randomly assigned to cisatracurium or placebo** groups - Double-blind design - Independent data and safety monitoring board
Generated Summary	- **Multicenter, double-blind trial** - **Randomly assigned patients to receive cisatracurium besylate or placebo.**
Score	Factual Correctness (Recall): **0.25** Factual Correctness (Precision): **0.67** Semantic Similarity : **0.86**

	Study Population
Ground Truth (ACURASYS)	The study enrolled **340 patients** from 20 ICUs in **France**. The patients were randomly assigned to either receive cisatracurium besylate (178 patients) or a placebo (162 patients) for 48 hours. The study population consisted of patients with severe ARDS, defined as a ratio of partial pressure of arterial oxygen (PaO2) to the fraction of inspired oxygen (FIO2) of less than 150, with a positive end-expiratory pressure of 5 cm or more of water, and a tidal volume of 6 to 8 ml per kilogram of predicted body weight.- Total Participants: 340 patients - Inclusion Criteria: - Receiving endotracheal **mechanical ventilation** - Acute hypoxemic respiratory failure - **Severe ARDS** (PaO2:FIO2 < 150) - ARDS onset within previous 48 hours - Cisatracurium Group: 178 patients - Placebo Group: 162 patients
Generated Summary	- **340 patients**
Score	Factual Correctness (Recall): **0.08** Factual Correctness (Precision): **1.0** Semantic Similarity : **0.45**

Wrapping up

- We developed a **knowledge-graph-based RAG model** able to **retrieve relevant document** regarding medical questions in critical care medicine (accuracy ~ 90%) and **provide paper summaries** to **facilitate clinical decision support.**

- **Next steps**
- Improve performance (retrieval, synthesis, reduce hallucination)
- Publish the work

Backup