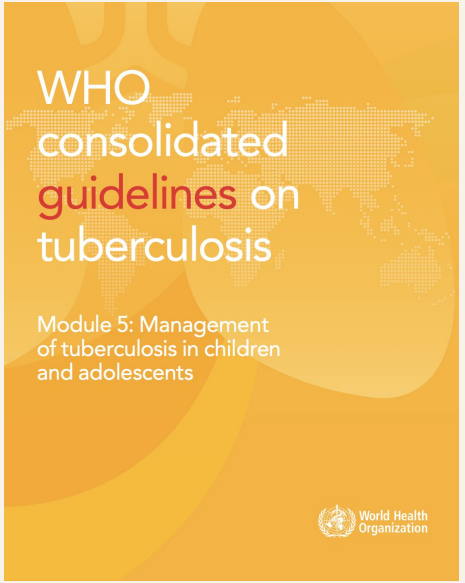
Retrieval-Augmented Generation for Medical Queries

Patrick Liu, Sophie Lin, Maria Arakelyan

Motivation

- Healthcare professionals struggle to access relevant information quickly due to volume of information
- LLMs offer potential, but they can hallucinate or provide inaccurate summaries
- Ensuring accuracy, trustworthiness, and up-to-date information in medical responses is crucial for patient safety and clinical decision-making
- Our goal: Use retrieval-augmented generation (RAG) to enhance the reliability of responses to medical queries by grounding them with trustworthy sources

Methodology



Dataset

- 6 Modules on Tuberculosis (TB) Guidelines from World Health Organization
    - Prevention, Screening, Diagnosis, Treatment, Management of tuberculosis in children and adolescents, Tuberculosis and Comorbidities
    - Ranges from ~100 - 400 pages
- Preprocessing
    - Manually scraped the documents for relevant sections
    - Removed Appendix, References, etc.
- Segmentation Strategy
    - Segmented based of section headers with max chunk size of 512 tokens
    - Split longer sections into multiple chunks with 50 tokens of overlap to ensure context across chunks

Methodology

Retrieval

- Query was encoded with Sentence-BERT embedding model
- Top-5 most semantically similar text chunks were retrieved using FAISS with cosine similarity
- Retrieved chunks were concatenated into a single context string with the query
    - "Given the following information to supplement your answer: {combined_texts}. Answer the question: {query}"
    - “Choices: {choices}. In your response, state the letter of the correct answer."

Methodology

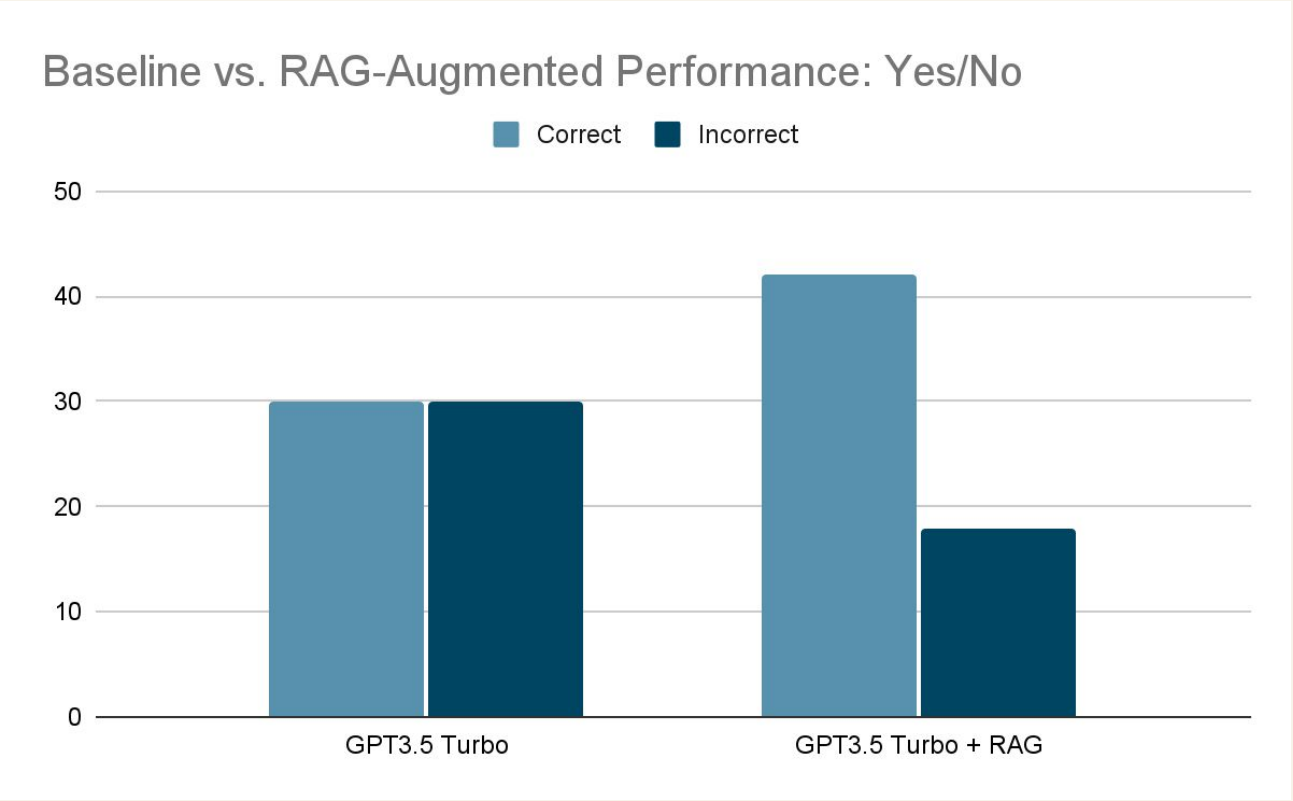
Evaluation Questions

- Generated by ChatGPT
- 3 Types:
  - 60 Yes/No
    - Ex. Should pregnancy disqualify women living with HIV from receiving isoniazid or rifampicin preventive treatment?
  - 57 Multiple Choice
    - Ex. Which of the following is NOT listed among the ethical principles emphasized for person-centred TBI care? Choices: [Informed consent, Non-coercion, Mandatory isolation, Confidentiality]
    - Used for Objective accuracy
  - 30 Free Response
    - Ex. Provide the recommended corticosteroid regimen for tuberculous meningitis.
      - LLM as a Judge - Claude 3.7 Sonnet

Results

Yes/No

Subset of answers were double checked directly with the text

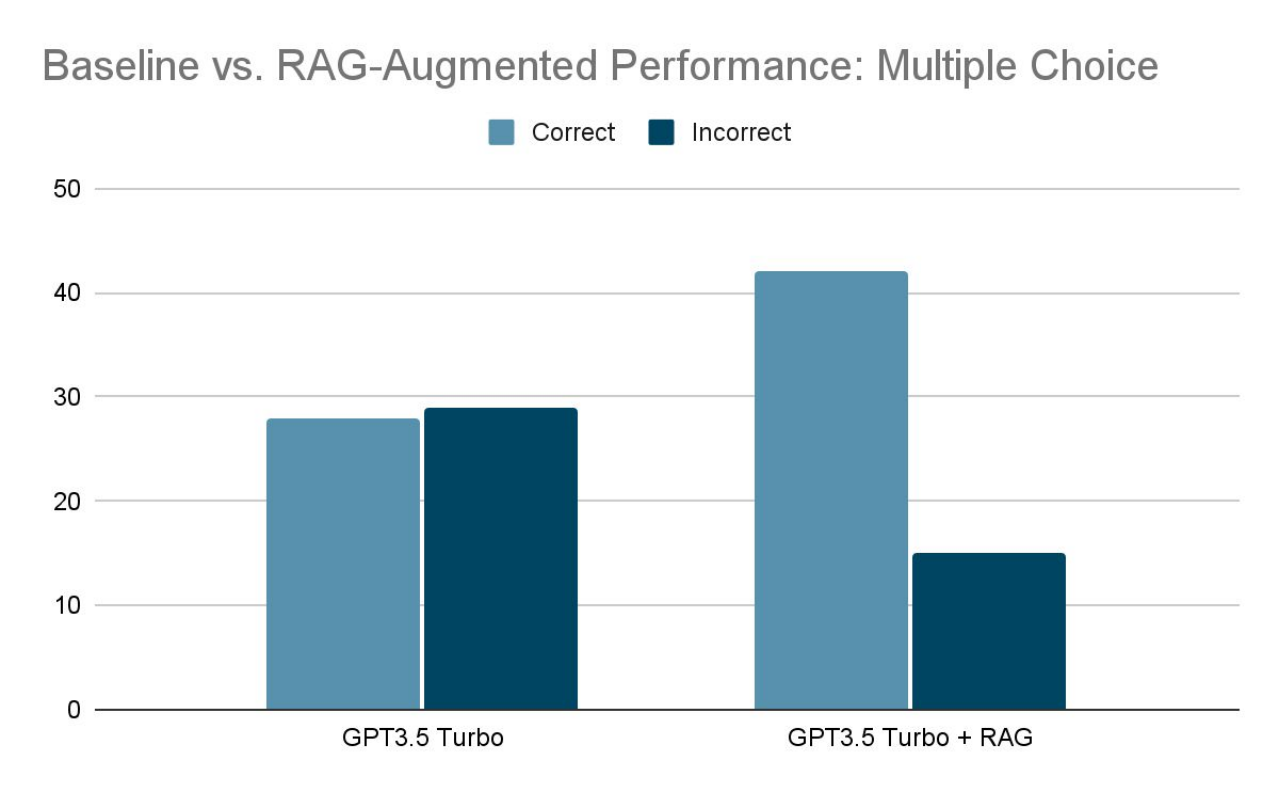


- Baseline achieved 30/60 correct answers
  - 50% accuracy
- RAG achieved 42/60 correct answers
  - improved to 70% accuracy

Results

Multiple Choice

Subset of answers were double checked with the text

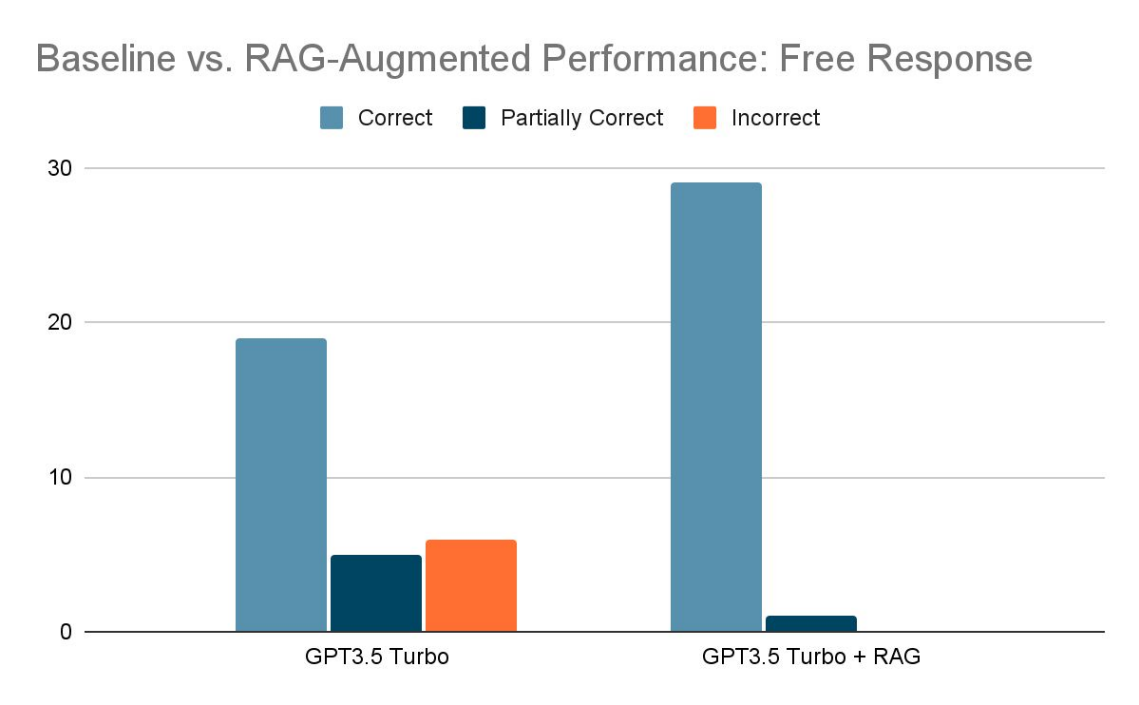


- Baseline achieved 28/57 correct answers
  - ~49.1% accuracy
- RAG achieved 42/57 correct answers
  - ~73.7% accuracy

Results

Free Response Example

Partially correct answers were defined by Claude as responses that were vague, included irrelevant information, or contained minor factual inaccuracies.



- Baseline achieved 19/30 correct answers
  - ~63.3% accuracy
- RAG achieved 29/30 correct answers
  - ~96.7% accuracy

Free Response Answer + Evaluation Example: gpt3.5 without Rag

Question 20: Identify three research priorities named for moderate-complexity automated NAATs.

1. Evaluating the performance of moderate-complexity automated NAATs in comparison to traditional laboratory methods for diagnosing infectious diseases.
2. Identifying and addressing potential challenges and limitations of moderate-complexity automated NAATs in different healthcare settings.
3. Assessing the cost-effectiveness and feasibility of implementing moderate-complexity automated NAATs in routine clinical practice for widespread use.

ANSWER CHECK: ⚠️ PARTIALLY CORRECT

- Lists three reasonable research priorities, but they are general and not specific to the document
- The priorities listed sound plausible but may not match the exact priorities specified in the source material
- Answer is logical but possibly not aligned with the specific priorities mentioned in the reference text

Free Response Example Excerpt: RAG

Should adults with HIV receive TB preventive treatment?

Yes, in most cases, adults with HIV should receive TB preventive treatment (TPT) if they do not have active tuberculosis (TB) disease and no contraindications to the preventive medications.

Key Points

1. **Why?**

People living with HIV are at much higher risk of developing active TB if exposed or infected with Mycobacterium tuberculosis. TB is a leading cause of illness and death in people with HIV.

2. **What is TB preventive treatment?**

TPT refers to medicines given to people at risk of TB (like those with HIV) to prevent the development of active TB. The most common regimens include isoniazid, rifapentine, or rifampicin for a defined period.

3. **Who should get it?**

All adults and adolescents living with HIV (including pregnant women and those on antiretroviral therapy), who do not have symptoms or signs of active TB, should be offered TPT according to World Health Organization (WHO) guidelines. Screening for active TB disease is important prior to initiating preventive treatment (using symptom screening, chest X-ray, or other tests).

4. **Who should NOT get it?**

Individuals with suspected or confirmed active TB, Those with contraindications to the drugs used (e.g., liver disease with isoniazid), A detailed assessment by a healthcare provider is necessary.

5. **Is testing for latent TB necessary?**

Testing for latent TB infection (e.g., tuberculin skin test or IGRA) is not required before starting TPT in people with HIV, according to WHO.

[This portion was cut off for space purposes]

References: WHO Guidelines on Tuberculosis Preventive Treatment (2020) link; CDC: Tuberculosis and HIV Co-Infection link

Always consult a healthcare provider for individualized risk assessment and management.

Results

Analysis: what it shows

- RAG-augmented prompts consistently outperform baseline prompting
  - Objective accuracy is higher
- Enhanced Quality in Free Response Answers
  - LLM-as-a-Judge evaluations indicate that RAG responses are more complete, accurate, and helpful than baseline
  - “The information you shared accurately reflects the WHO and CDC guidance” (taken from Claude’s evaluation of our free response)
- Benefit of Contextual Retrieval
  - Overlapping chunking and document curation likely improved relevance of retrieved passages

Results

Analysis: what it shows

RAG changed the right answer to a wrong answer for a few questions (around 3 for multiple choice)

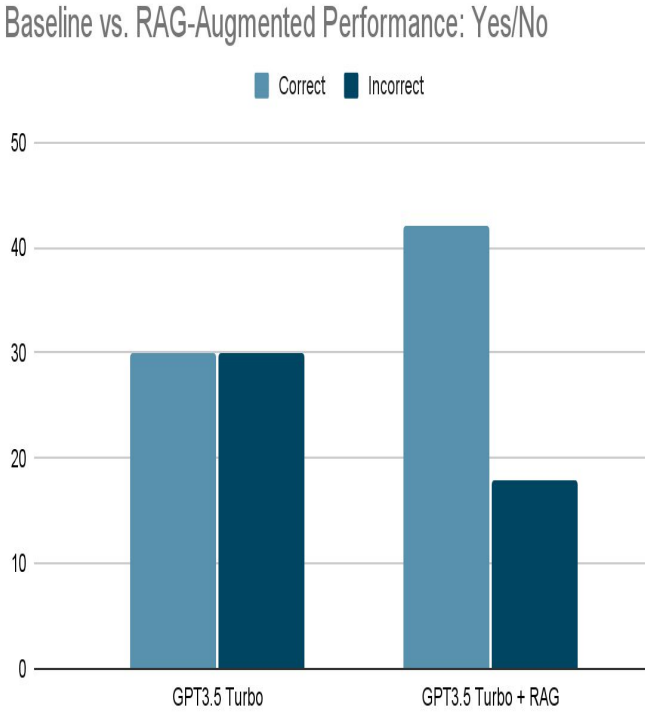
- Potential Reasons:
  - Irrelevant or Noisy Chunks
    - Retrieved chunks may not be relevant, causing the model to generate incorrect answers
  - Hallucinations from Retrieved Chunk
    - The model may hallucinate details from retrieved passages, through misinterpreting ambiguous phrasing or inferring incorrect relationships between facts
  - Chunk Overlap Issues
    - Insufficient sized chunk or overlap might miss critical context
  - Inadequate Retrieval Scope
    - Retrieving too few chunks may not provide enough context for the correct answer.

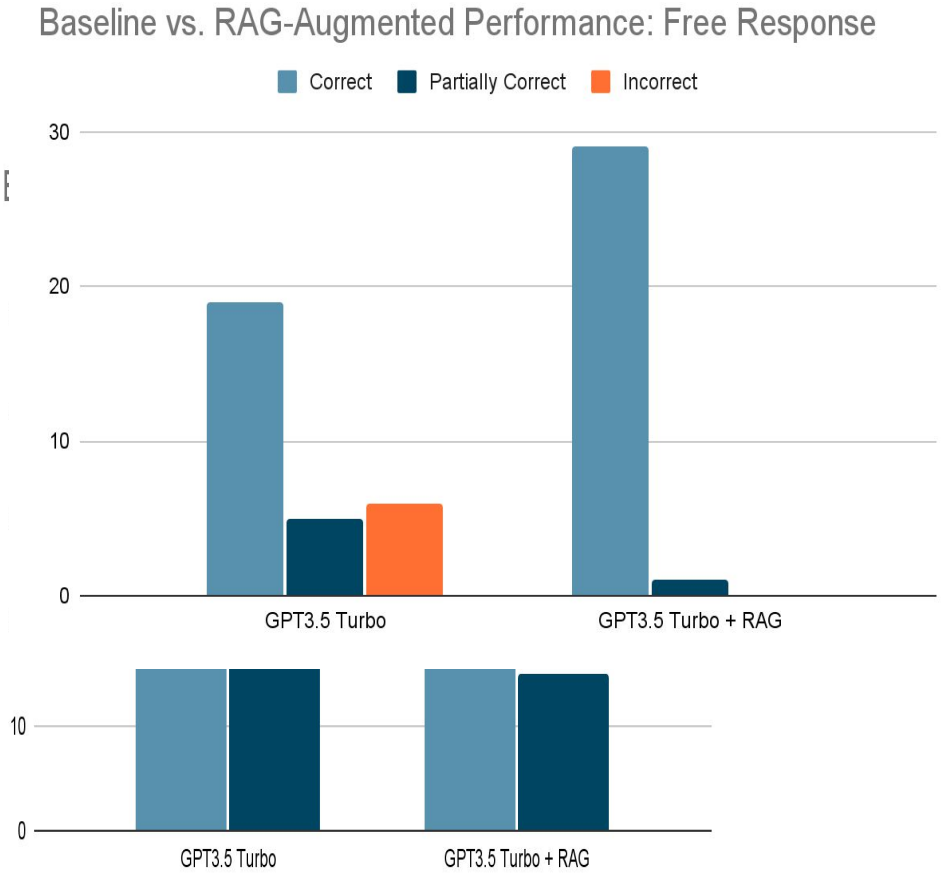
Results

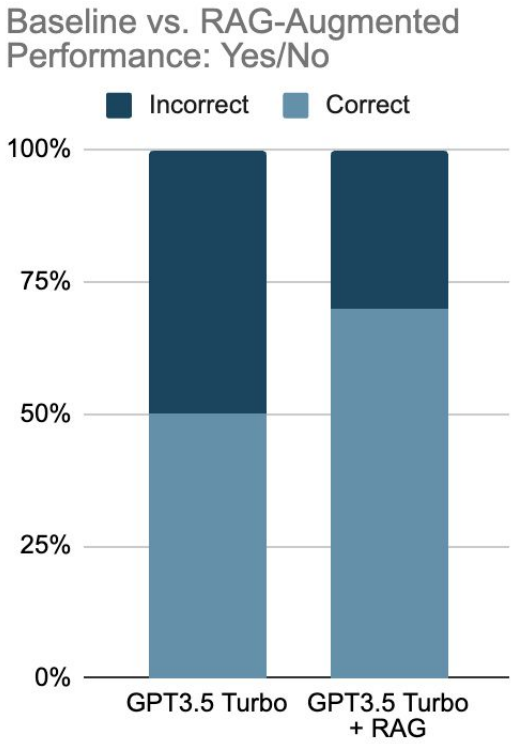
Analysis: what it doesn’t show

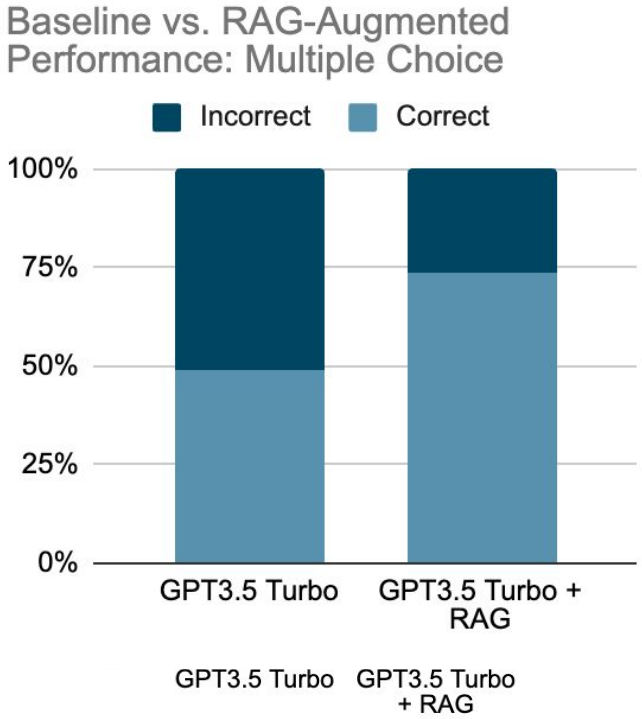
- Questions generated directly based on the document
  - Model may "perform well" simply because it's being tested on concepts that were designed to be aligned with the source
  - Inherent biases in these documents
- Ratings rely on another LLM's subjective judgment (for LLM as a judge)
  - LLMs tend to prefer longer, more detailed responses
  - LLM as a judge can hallucinate
- Language and Framing
  - More technical jargon can influence how an LLM interprets answers
- Doesn’t account for response latency, readability, or user experience

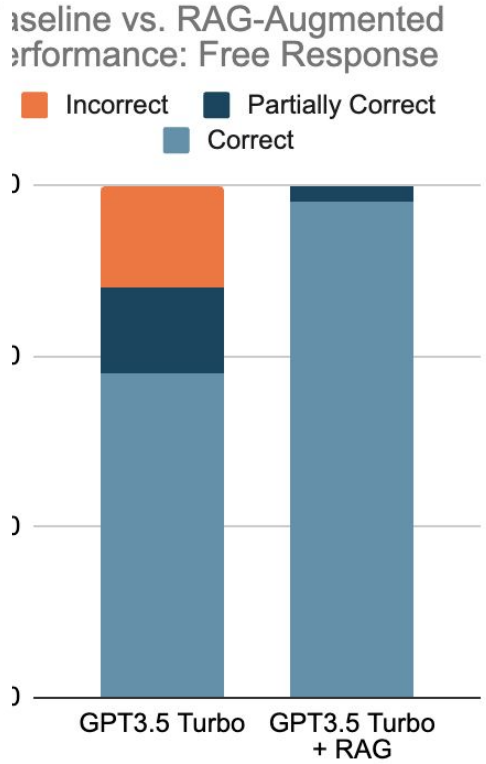
Thank you for listening!

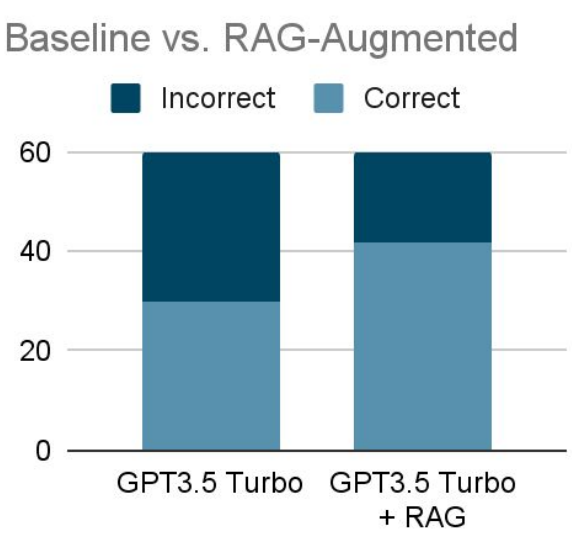

Baseline vs. RAG-Augmented Performance: Yes/No
Correct
Incorrect
50
40
30
20
10
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG

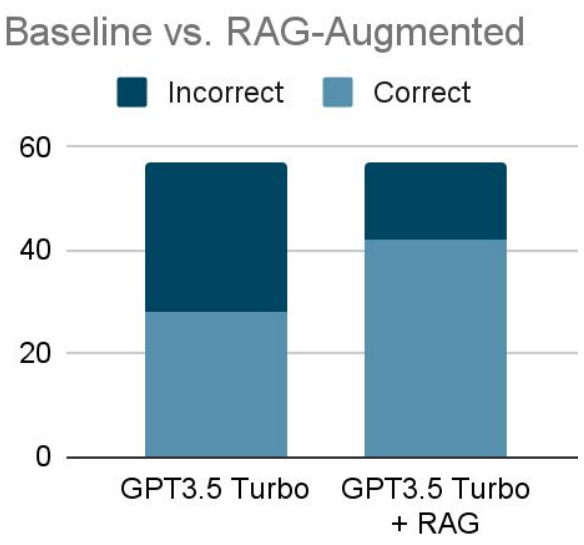

Baseline vs. RAG-Augmented Performance: Free Response
Correct
Partially Correct
Incorrect
30
20
10
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG
10
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG

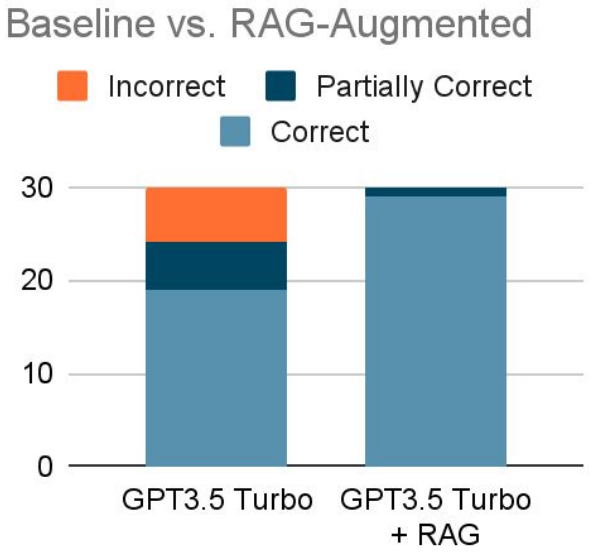

Baseline vs. RAG-Augmented Performance: Yes/No
Incorrect
Correct
100%
75%
50%
25%
0%
GPT3.5 Turbo
GPT3.5 Turbo + RAG


Baseline vs. RAG-Augmented Performance: Multiple Choice
Incorrect
Correct
100%
75%
50%
25%
0%
GPT3.5 Turbo
GPT3.5 Turbo + RAG
GPT3.5 Turbo
GPT3.5 Turbo + RAG


aseline vs. RAG-Augmented
erformance: Free Response
Incorrect
Partially Correct
Correct
GPT3.5 Turbo
GPT3.5 Turbo + RAG


Baseline vs. RAG-Augmented
Incorrect
Correct
60
40
20
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG


Baseline vs. RAG-Augmented
Incorrect
Correct
60
40
20
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG


Baseline vs. RAG-Augmented
Incorrect
Partially Correct
Correct
30
20
10
0
GPT3.5 Turbo
GPT3.5 Turbo + RAG