



Retrieval Augmented Generation (RAG) for Biomedicine

Presenter: Qiao Jin, M.D. (@DrQiaoJin)

Yale



National Library of Medicine
National Center for Biotechnology Information

Large Language Models **Hallucinate**

Q1

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source article title and PMID.

Large Language Models **Hallucinate**

Q1

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source article title and PMID.



Sure, here are some of the most common symptoms associated with long-term sequelae of COVID-19, along with their corresponding source article title and PMID:

1. Fatigue and Muscle Weakness: "Post-acute COVID-19 syndrome: a case series and multidisciplinary approach to treatment" (PMID: 33571309)
2. Shortness of Breath: "Symptomatic Outcomes in Young Adults With COVID-19" (PMID: 32936950)
3. Joint Pain: "Musculoskeletal symptoms in COVID-19 patients" (PMID: 32767232)
4. Chest Pain: "Chest CT Findings in Long-COVID Patients with Dyspnea and Chest Pain" (PMID: 33828099)



**Looking
Plausible**

Large Language Models **Hallucinate**

Q1

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source article title and PMID.

fabricated




Sure, here are some of the most common symptoms associated with long-term sequelae of COVID-19, along with their corresponding source article title and PMID:


1. Fatigue and Muscle Weakness: "Post-acute COVID-19 syndrome: a case series and multidisciplinary approach to treatment" (PMID: 33571309)
2. Shortness of Breath: "Symptomatic Outcomes in Young Adults With COVID-19" (PMID: 32936950)
3. Joint Pain: "Musculoskeletal symptoms in COVID-19 patients" (PMID: 32767232)
4. Chest Pain: "Chest CT Findings in Long-COVID Patients with Dyspnea and Chest Pain" (PMID: 33828099)




Retrieval as a Solution to Hallucination

 Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source PMID.

Providing Relevant Materials

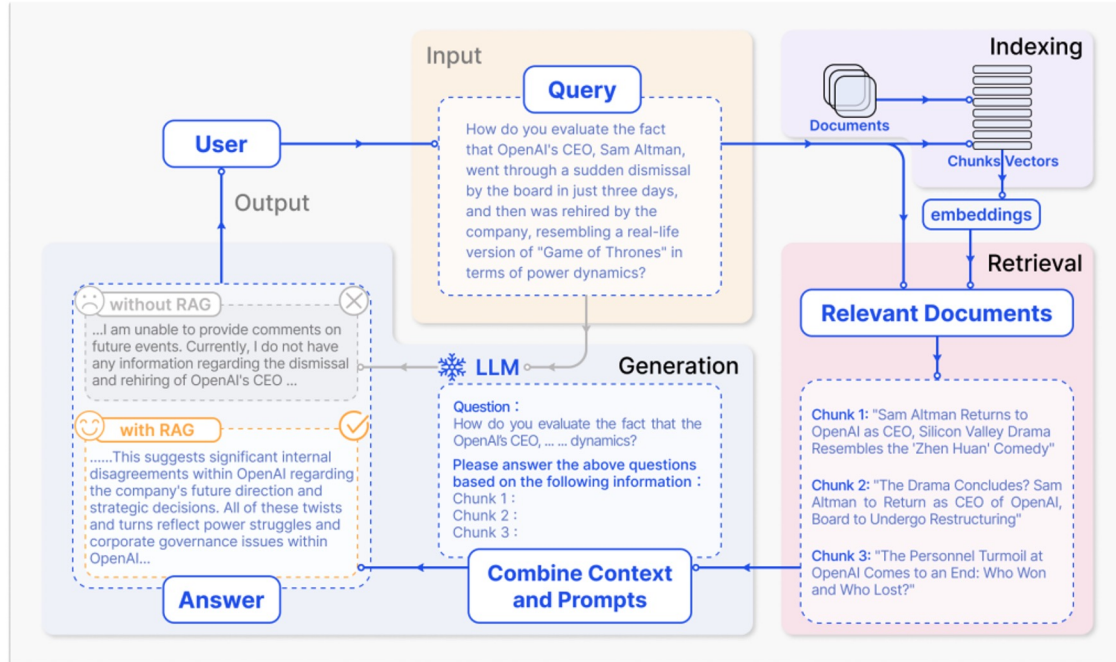


 Based on the sources provided, some common symptoms associated with the long-term sequelae of COVID-19 include:

1. Fatigue and malaise (PMID: 35105985; PMID: 36215063)
2. Dyspnea (difficulty breathing) (PMID: 35105985; PMID: 36215063)
3. Defects in memory and concentration (PMID: 35105985)
4. Coanitive problems (PMID: 36215063; PMID: 35934007)

Much Better

Retrieval-augmented Generation (RAG) Architecture



Medical Information Retrieval-Augmented Generation Evaluation






| 5 Datasets | 7,663 Questions | 2-4 Choices |
|--|---|------------------|
|  MMLU-Med | Which of the following best describes ... ? | A / B / C / D |
|  MedQA-US | A 72-year-old man comes to the physicians ... ? | A / B / C / D |
|  MedMCQA | Axonal transport is: | A / B / C / D |
|  PubMedQA* | Is anorectal endosonography valuable ... ? | Yes / No / Maybe |
|  BioASQ-Y/N | Is medical hydrology the same as Spa ... ? | Yes / No |

Figure 1: Composition of the MIRAGE benchmark.

MedRAG Toolkit

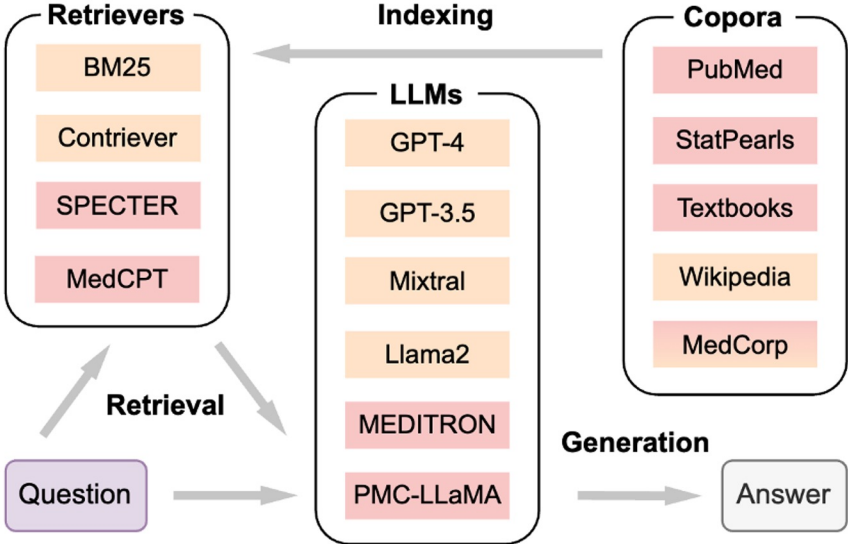
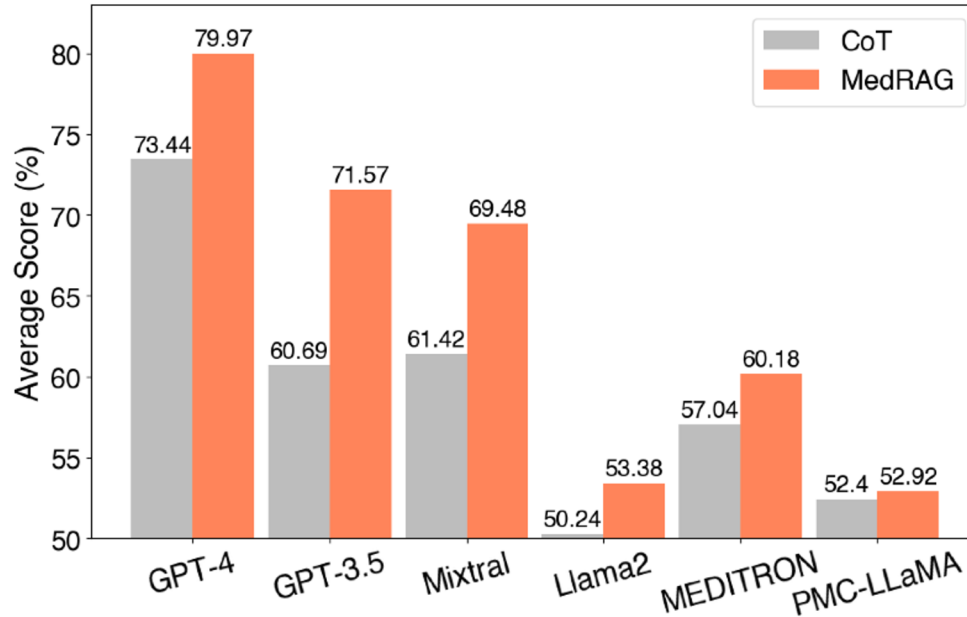


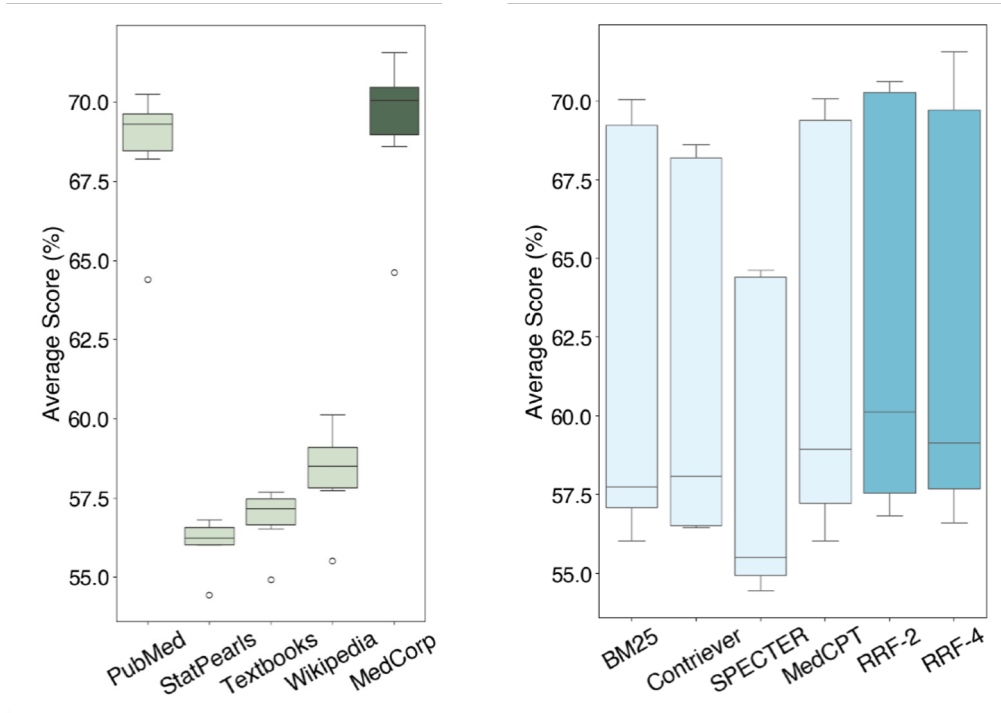
Figure 2: Component overview of the MEDRAG toolkit.

MedRAG Improves Medical QA Performance

On MIRAGE, MedRAG improves the accuracy of different LLMs by **up to 18%**



Effects of Corpora and Retrievers in MedRAG



(left) Corpora have more influence on performance

(right) Different retrievers get less different results

(overall) The combination of various medical corpora and retrievers achieves the best performance

MedRAG: Scaling Properties

Log-linear improvement with # retrieved snippets

Favoring long-context LLMs

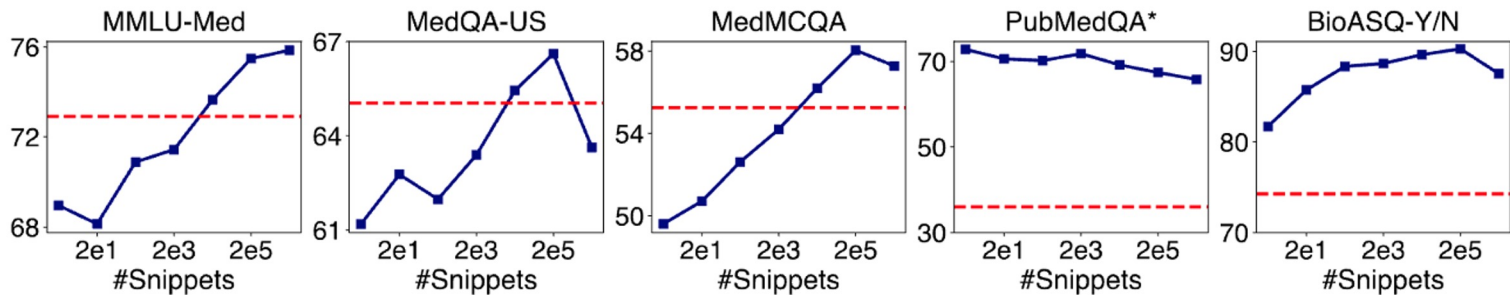
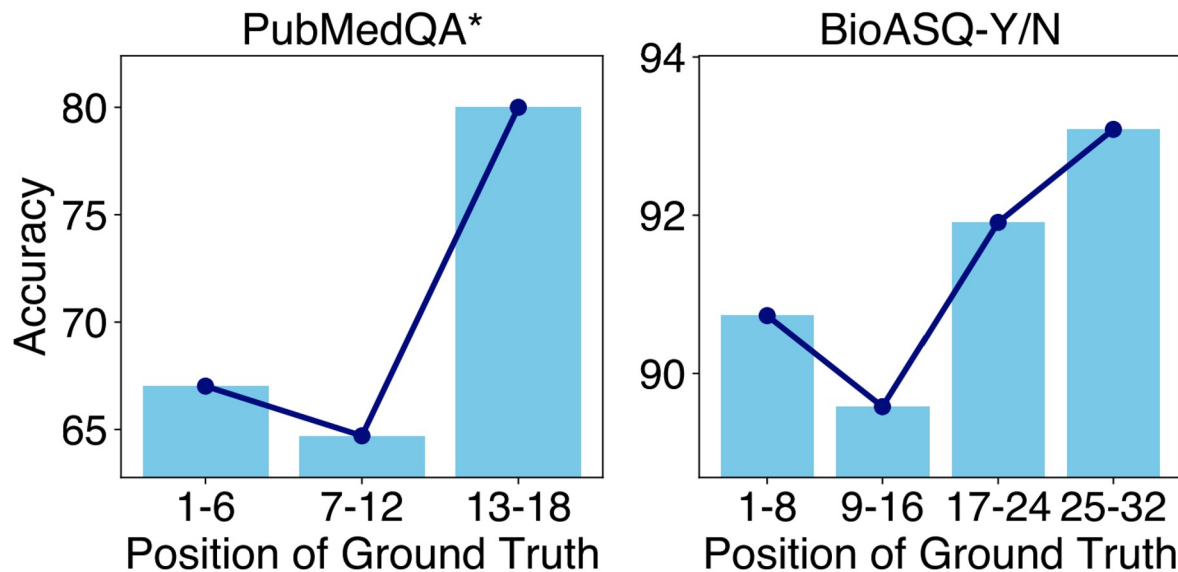


Figure 3: MEDRAG accuracy with different numbers of retrieved snippets. Red dotted lines denote CoT performance.

“Lost-in-the-middle”

MedRAG performance is lowest when relevant article is in the middle of prompt



Coding & QA


https://colab.research.google.com/drive/1HoYWH0yJBQliWC_XRKL-COwfdbXTCDUj?usp=sharing

RAG is NOT Enough

🌐 Model: Web Browsing

Q1

What's the gene associated with SNP rs1241371358?



🕒 Finished browsing▼

I'm sorry, but I was unable to find information about the gene associated with the SNP rs1241371358. It seems that there is limited or no public information available for this specific SNP. This might be due to the SNP being relatively unknown, or not having been characterized in any published research to date.

If you have more specific information about the context in which this SNP is being studied or mentioned, I might be able to provide more help. Alternatively, if you're working in a research setting, you might want to contact a geneticist or use a specific genetic database tool for further information.

14

GeneGPT: First Tool-augmented LLM in Biomedicine



×



Teaching LLMs to
use NCBI Web APIs

GeneHop question (SNP gene function):

What is the function of the gene associated with SNP rs1241371358? Let's decompose the question to sub-questions and solve them step by step.

GeneGPT output:

Sub-question 1: Which gene is SNP rs1241371358 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=10&retmode=json&id=1241371358]
```

->[API call results]

Answer: LRRC23

Sub-question 2: What is the function of LRRC23?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&retmax=10&retmode=json&id=10233]
```

->[API call results]

Answer: Predicted to be active in cytosol. (correct)

GeneGPT: First Tool-augmented LLM in Biomedicine

| GeneTuring task | GPT-2 | BioGPT | BioMedLM | GPT-3 | ChatGPT | New Bing | GeneGPT (ours) | | Med-Gemini |
|----------------------------|-------|-------------|----------|-------|---------|-------------|----------------|---------------|-------------|
| | | | | | | | -full | -slim | |
| Nomenclature | | | | | | | | | |
| Gene alias | 0.00 | 0.00 | 0.04 | 0.09 | 0.07 | 0.66 | <u>0.80</u> * | 0.84 * | 0.72 |
| Gene name conversion | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | <u>0.85</u> | 1.00 | 1.00 | 1.00 |
| Average | 0.00 | 0.00 | 0.02 | 0.05 | 0.04 | 0.76 | <u>0.90</u> | 0.92 | 0.86 |
| Genomic location | | | | | | | | | |
| Gene SNP association | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 * | 1.00 | 0.00 |
| Gene location | 0.01 | 0.04 | 0.12 | 0.09 | 0.09 | 0.61 | <u>0.62</u> | 0.66 | 0.83 |
| SNP location | 0.03 | <u>0.05</u> | 0.01 | 0.02 | 0.05 | 0.01 | 1.00 | <u>0.98</u> | 0.00 |
| Average | 0.01 | 0.03 | 0.04 | 0.04 | 0.05 | 0.21 | <u>0.87</u> | 0.88 | 0.28 |
| Functional analysis | | | | | | | | | |
| Gene disease association | 0.00 | 0.02 | 0.16 | 0.34 | 0.31 | 0.84 | <u>0.76</u> * | 0.66 | 0.82 |
| Protein-coding genes | 0.00 | 0.18 | 0.37 | 0.70 | 0.54 | <u>0.97</u> | 0.76 | 1.00 | 1.00 |
| Average | 0.00 | 0.10 | 0.27 | 0.52 | 0.43 | 0.91 | 0.76 | <u>0.84</u> | 0.91 |
| Sequence alignment | | | | | | | | | |
| DNA to human genome | 0.02 | <u>0.07</u> | 0.03 | 0.00 | 0.00 | 0.00 | 0.44 * | 0.44 * | 0.00 |
| DNA to multiple species | 0.02 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | <u>0.86</u> | 0.88 | 0.13 |
| Average | 0.02 | 0.04 | 0.02 | 0.10 | 0.00 | 0.00 | <u>0.65</u> | 0.66 | 0.07 |
| Overall average | 0.00 | 0.04 | 0.08 | 0.16 | 0.12 | 0.44 | <u>0.80</u> | 0.83 | 0.50 |



Med-Gemini

0.72
1.00
0.86
0.00
0.83
0.00
0.28
0.82
1.00
0.91
0.00
0.13
0.07
0.50

Thank you!

Questions?