



INSTITUTE FOR DEFENSE ANALYSES

**Talk to the NDAA: Evaluating Retrieval
Augmented Generation (RAG) on
Congressional Legislation using Large
Language Models (LLM)**

Kevin Garrison, Project Leader

William M. Fisher

Astrid I. Berge

August 2024

Distribution Statement A.
Approved for public release:
distribution is unlimited.

IDA Product 3001199

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305



The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, Project C5240, “Generative AI Use Cases,” for the IDA. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgements

Arun S. Maiya

For More Information

Kevin Garrison, Project Leader
kgarriso@ida.org, 703-933-6545

Margaret E. Myers, Director, Information Technology and Systems Division
mmyers@ida.org, 703-578-2782

Copyright Notice

© 2024 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Executive Summary

One of the recurring use cases for large language models (LLMs) is the generation of specific answers on text data outside of the model training set. Retrieval augmented generation (RAG) systems augment LLMs by allowing them to retrieve information from sources outside of those they were trained on. However, it is unclear how different questions and parameter settings affect the quality of answers produced by these systems. In this analysis, we created a RAG system to retrieve information from the 2024 NDAA in response to questions. We tested the system with a set of eight questions, varied the number of chunks of text and the size of the chunks retrieved in response to each question, and noted how these affected response quality. We used GPT-4 and Mistral-7b as the base LLMs and evaluated their responses using the LlamaIndex Faithfulness and Relevancy evaluators, which determine whether a RAG system's response is faithful to the text retrieved and relevant to the question asked. We evaluated the scores across the different parameter settings and questions and compared how each LLM passed or failed the Faithfulness and Relevancy metrics. We found that although there were small performance improvements when retrieving smaller chunks of text in response to questions, the LLM used and the complexity of the question were the most important factors in RAG performance. Specifically, GPT-4 generally performed better than Mistral-7b, and both struggled to answer questions that were not directly addressed in the NDAA text.

Contents

1.	Introduction	1-1
	A. Methods	1-1
	B. Response Evaluation	1-3
2.	Response Comparison	2-1
	A. Example 1	2-1
	1. GPT-4	2-1
	2. Mistral-7b	2-1
	3. Analysis	2-2
	B. Example 2	2-2
	1. GPT-4	2-2
	2. Mistral-7b	2-2
	3. Analysis	2-3
	C. Example 3	2-3
	1. GPT-4	2-3
	2. Mistral-7b	2-4
	3. Analysis	2-4
	D. Similarity Scores	2-4
	1. What is said about amphibious vehicles?	2-6
	2. Tell me about artificial intelligence in the Coast Guard.	2-7
	3. What sections mention the military health system?	2-7
	E. Limitations.....	2-7
	F. Conclusions	2-8

1. Introduction

One of the recurring use cases for large language models (LLMs) is the generation of specific answers on text data outside of the model training set. Retrieval augmented generation (RAG) systems augment regular LLMs by retrieving the most relevant parts of input text data in response to a prompt. RAG is best employed in cases where there is data outside the original set of general-purpose text that an LLM is pre-trained on and especially in cases when prompts exceed context lengths. This also reduces the likelihood of hallucinations and removes the need to pre-train on sensitive information. While RAG systems have become increasingly popular, the different parameters mean that there is no consensus on the best way to set one up. Using the NDAA as an example, researchers could use a RAG LLM system to determine DoD’s existing priorities and potential projects for the Institute for Defense Analyses (IDA). In this analysis, we examine how changing the parameters and types of prompts for a RAG LLM system using text from the NDAA changes the output for a given model.

A. Methods

We used the RAG technique with the draft of the 2024 NDAA from the U.S. Senate to test various queries under different sets of parameters. We used Python to develop the RAG system; more specifically, we used various packages from the Python release of LlamaIndex,¹ an open source framework commonly used to build LLM applications. Within LlamaIndex, we used the SentenceSplitter package to split the text into chunks and the all-MiniLM-L6-v2 sentence and short paragraph transformer model implemented in HuggingFace Embeddings for text embedding. Together with the split documents and the embeddings, we used the Facebook AI Similarity Score (FAISS) library from Meta to make an index of the chunks of documents that could be queried using the VectorStoreIndex module within LlamaIndex. Finally, we used the VectorStoreIndex as_query_engine method for document retrieval in response to prompts.

We used Mistral-7b and GPT-4 as the LLMs for this work. Designed by Mistral AI, Mistral-7b is a 7.3 billion parameter model designed to outperform other major downloadable LLMs, such as Meta’s Llama-2, on most benchmarks.² GPT-4 is the latest model released from OpenAI, with over 1 trillion parameters.³ Unlike Mistral-7b, GPT-4 cannot be run locally and instead must be used through the OpenAI API. This means that although Mistral-7b was running locally on a cluster of CPUs locally, GPT-4 was running on cloud resources.

¹ <https://pypi.org/project/llama-index/>

² <https://mistral.ai/news/announcing-mistral-7b/>

³ <https://www.semafor.com/article/11/01/2023/microsoft-pushes-the-boundaries-of-small-ai-models>

The prompts were given in the following format:

```
“Context information is below.\n”  
“-----\n”  
“{context_str}\n”  
“-----\n”  
“Given the context information and not prior knowledge, ”  
“answer the query.\n”  
“Query: {query_str}\n”  
“Answer:”
```

where {context_str} is replaced by the documents retrieved and {query_str} is replaced by the question inputted into the retriever.

We varied two hyperparameters for this exercise. The first was the size of the chunks that the text was split into. We tested chunk sizes of 128, 256, 512, and 1,024 tokens (i.e., the unit of text that an LLM processes and generates). On average, a token is about four characters long. This does not mean that every chunk was exactly the length specified in the parameter; the SentenceSplitter tries to keep sentences and paragraphs together unless it encroaches on the chunk size limit. Additionally, chunks cannot cross sections of the NDAA, so some chunks toward the end of sections were under the parameter limit. We tested different chunk sizes to determine if additional context improved the model’s performance. In fact, some of our results show that more context does not always improve performance. We also varied the number of chunks retrieved in conjunction with chunk size (from 2 to 7).

We asked the model eight questions under each combination of chunk size and number of documents retrieved, which resulted in 24 responses.

1. Tell me about artificial intelligence in the Coast Guard.
2. What are key concerns for the Space Force?
3. What are some concerns for the Space Force that outside government contractors may be needed to address?
4. What is said about amphibious vehicles?
5. What sections mention the military health system?
6. What studies involving recruiting and retention are in the NDAA?
7. What work in Taiwan is being funded?
8. Which study involving recruiting and retention is the highest priority for a federally funded research and development center (FFRDC)? Explain why.

To determine the efficiency of Mistral-7b and GPT-4 under each combination of parameters, we recorded each LLM’s response time. We used LlamaIndex’s Faithfulness and Relevancy evaluator packages to measure the quality of the responses. Both of these evaluation packages use an LLM (in this case, GPT-4) to determine the quality of a response. The Faithfulness evaluator compares the generated answer to the retrieved context information to determine if the LLM’s answer is hallucinatory or not.⁴ The Relevancy evaluator compares the response and context information to the question to determine if the question was successfully answered.⁵ Both of these evaluators output a “pass” or “fail” for each question, context, and response set that they evaluate.

B. Response Evaluation

Figure 1-1 displays mean response times from Mistral-7b and GPT-4 under different parameter settings. While both show an increase in response time as the number of documents retrieved and the chunk size increase, Mistral-7b’s response time is consistently longer than GPT-4’s. This makes sense considering GPT-4’s cloud computing resources for completing parallel tasks, which results in a greater rate of tokens generated per second despite GPT-4’s greater complexity. This difference is present at the setting where the least amount of context is provided (two documents retrieved and a chunk size of 128 tokens), as Mistral-7b takes nearly four times as long to respond as GPT-4. However, the difference in response times increases as the amount of context provided increases, with GPT-4 responding over 20 times faster with seven documents retrieved and a chunk size of 1,024 tokens.

⁴ <https://ts.llamaindex.ai/modules/evaluation/modules/faithfulness>

⁵ https://docs.llamaindex.ai/en/stable/examples/evaluation/relevancy_eval.html

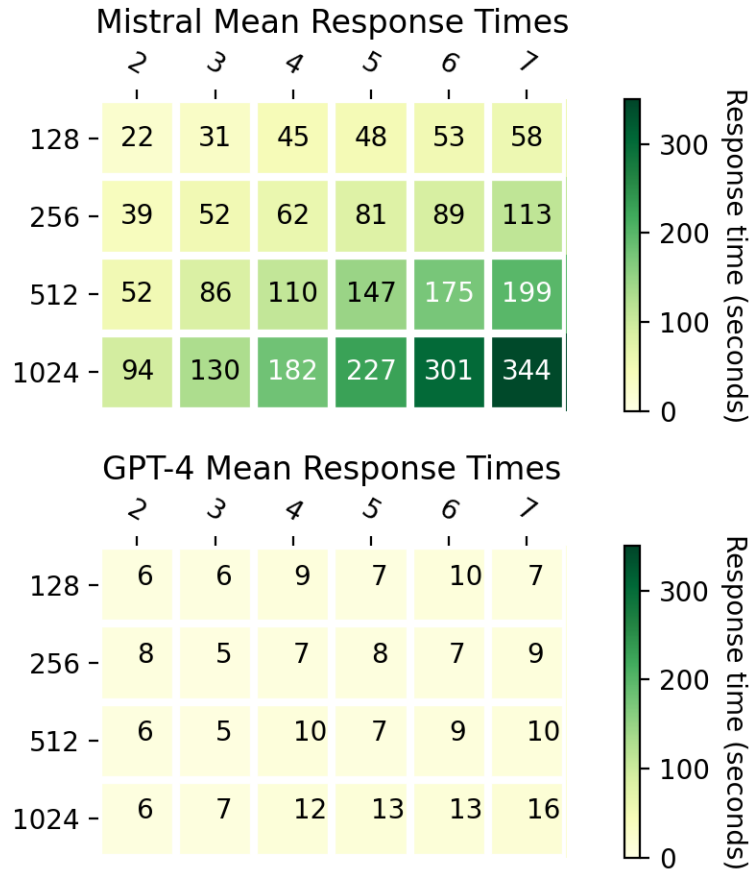


Figure 1-1. Mistral-7b and GPT-4 Mean Response Times by Chunk Size and Number of Retrieved Documents

Overall, Mistral-7b outperformed GPT-4 on the Faithfulness metric, receiving a pass result 89% of the time compared to GPT-4’s 77%. The pass/fail results for the LLMs on each prompt are shown in Figures 1-2 and 1-3. Although both did well on the metric overall, each LLM had particular questions for which its failure rate was higher than average. In particular, Mistral-7b failed 29% of the time on the Prompt 8, “What are some concerns for the Space Force that outside government contractors may be needed to address?” This is one of the more complex questions, and the retrieved documents retrieved may not have a direct answer to it, which may have led the LLM to respond in a way that strays from the documents. GPT-4 failed 58% of the time on Prompt 4, “What is said about amphibious vehicles?” The NDAA does not explicitly mention amphibious vehicles but it does mention amphibious warships, which GPT-4 did mention in its responses. However, Mistral-7b also mentioned amphibious warships in its Prompt 4 responses and had a higher pass rate on the prompt, so it is unclear why the two LLMs received different results from the metric. We discuss the Mistral-7b and GPT-4 responses further in the next section.

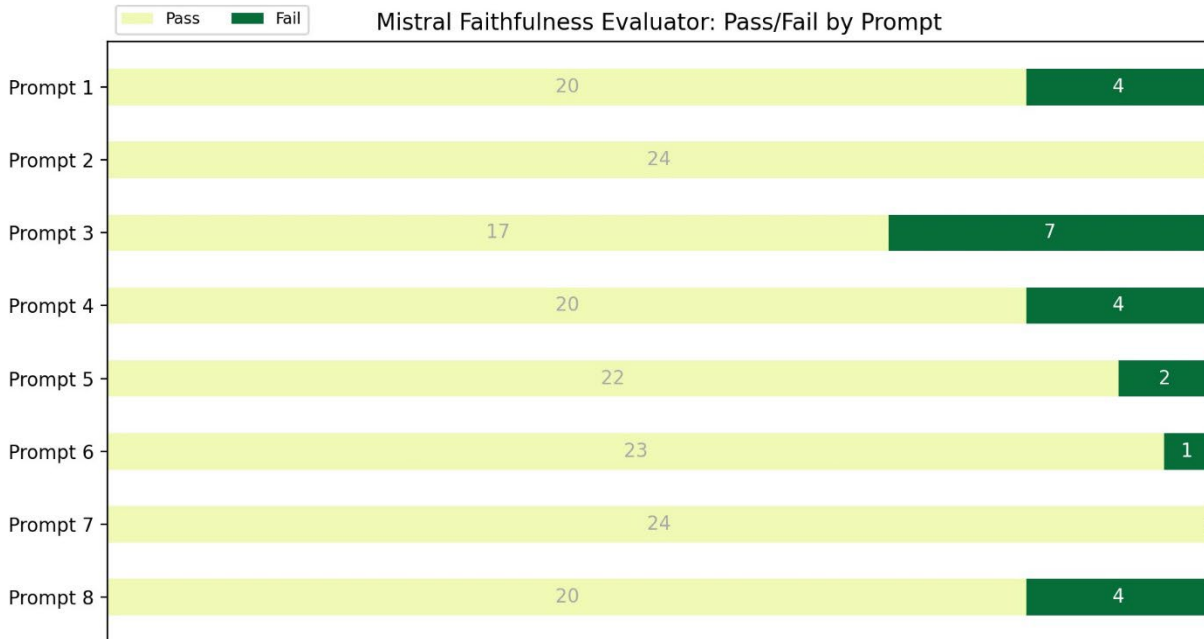


Figure 1-2. Mistral-7b Faithfulness Indicator Results

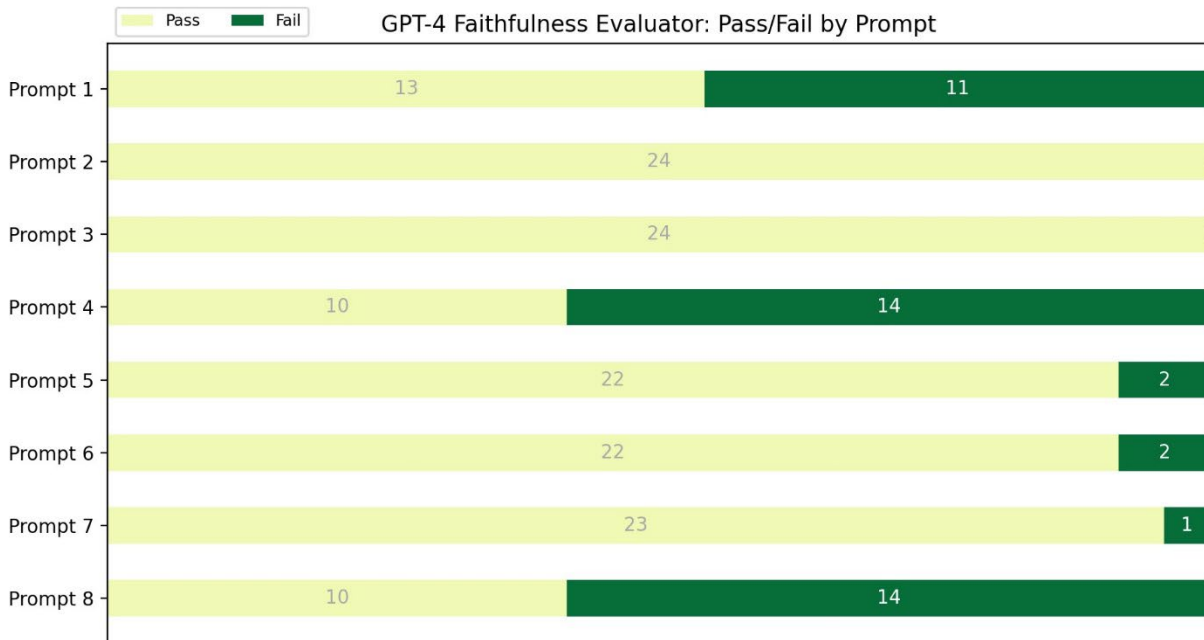


Figure 1-3. GPT-4 Faithfulness Indicator Results

GPT-4 outperformed Mistral-7b on the Relevancy metric, receiving a pass result 73% of the time compared to Mistral-7b’s 45%. The pass/fail results for the LLMs on each prompt are shown in Figures 1-4 and 1-5. GPT-4 performed worst (58% fail rate) on Prompt 1, “Tell me about artificial intelligence in the Coast Guard.” This is likely because the NDAA does not mention

artificial intelligence in the Coast Guard, so GPT-4 did not directly answer the question in some of the responses, and in others it attempted to relate AI use in other parts of the NDAA to the Coast Guard. Mistral-7b performed worst (67% failure rate) on the Prompt 6, “What studies involving recruiting and retention are in the NDAA?.” Mistral-7b often used NDAA text that generally referenced recruiting and retention but did not refer to the studies on those issues, which is likely why it failed the Relevancy metric to such a degree.

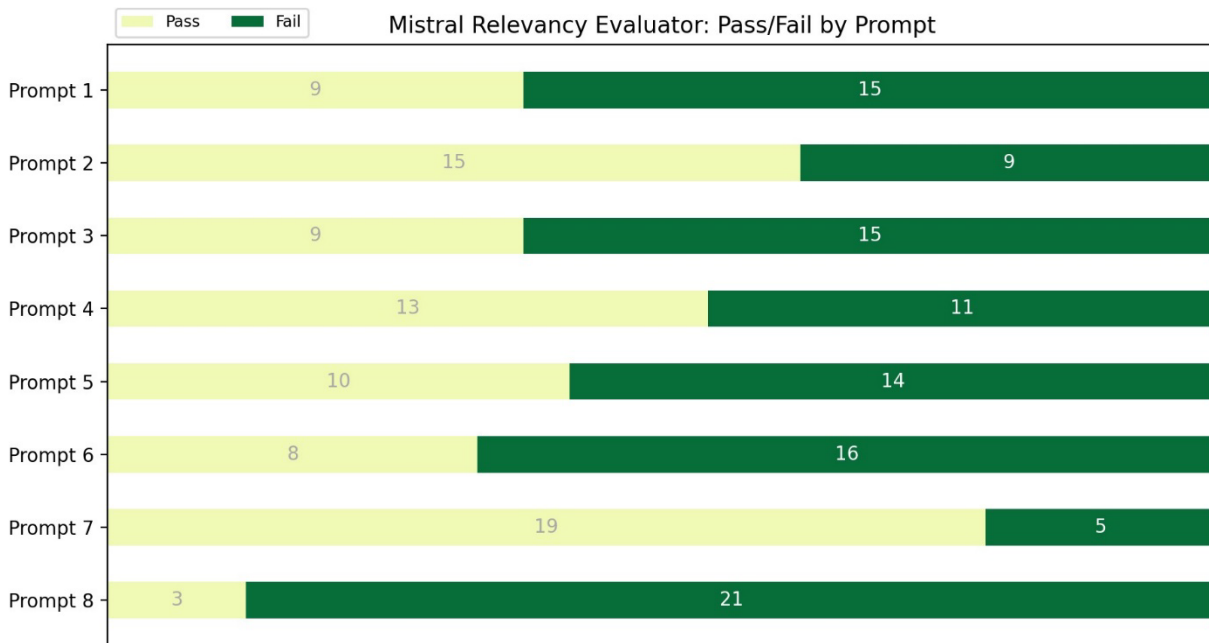


Figure 1-4. Mistral-7b Relevancy Indicator Results

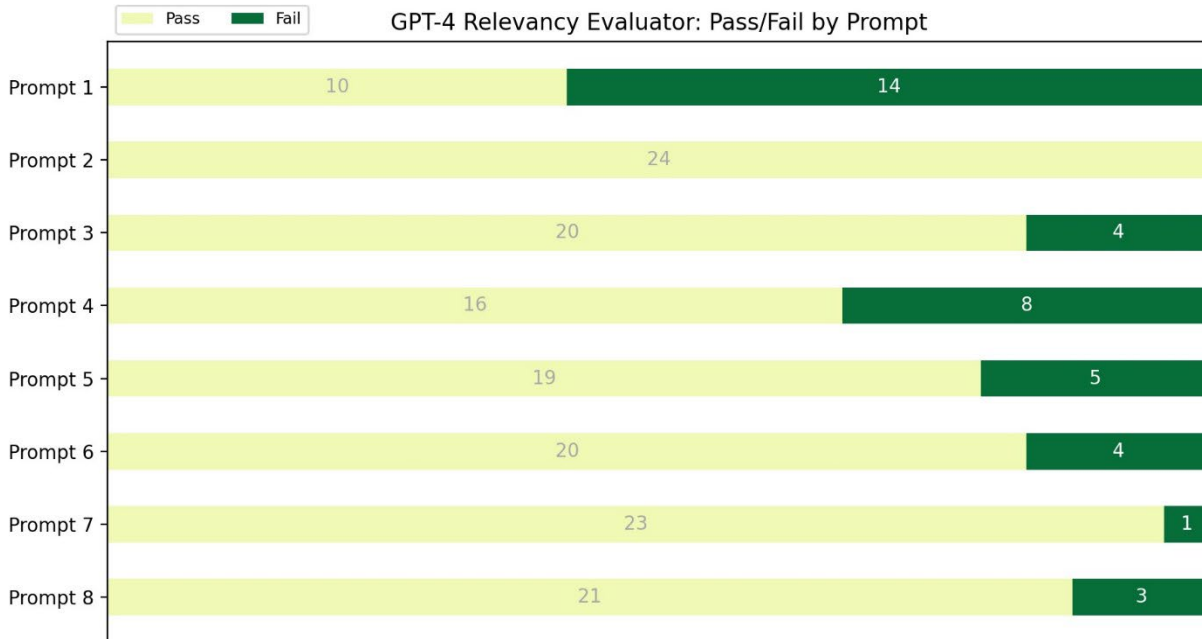


Figure 1-5. GPT-4 Relevancy Indicator Results

Figure 1-6 and Figure 1-7 display the passing rates of Mistral-7b and GPT-4 for the Relevancy and Faithfulness metrics, respectively, based on the chunk size and number of documents retrieved. Mistral-7b and GPT-4 both received high scores from the Relevance metric when using smaller chunk sizes, but there are no other clear patterns on how the number of documents retrieved or chunk size affects the quality of responses. This may mean that the most important factors determining the quality of a RAG system’s response to a prompt are the prompt’s complexity and availability of the information requested, as opposed to the quantity of information being retrieved.

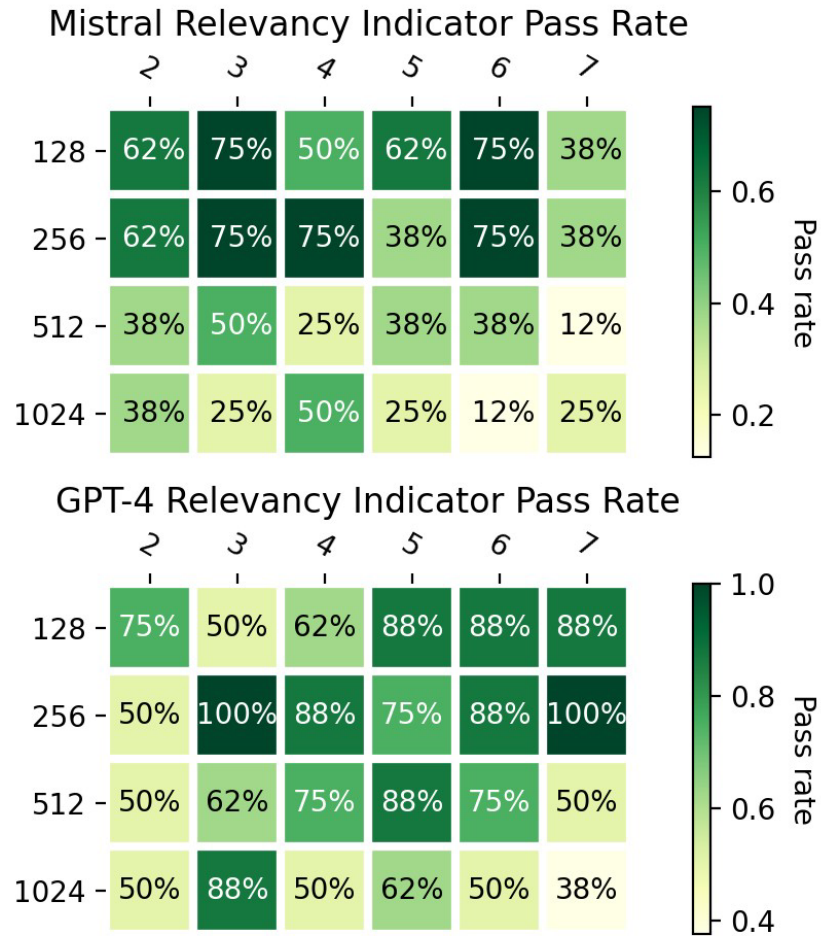


Figure 1-6. Mistral-7b and GPT-4 Relevancy Passing Rates by Chunk Size and Number of Documents Retrieved

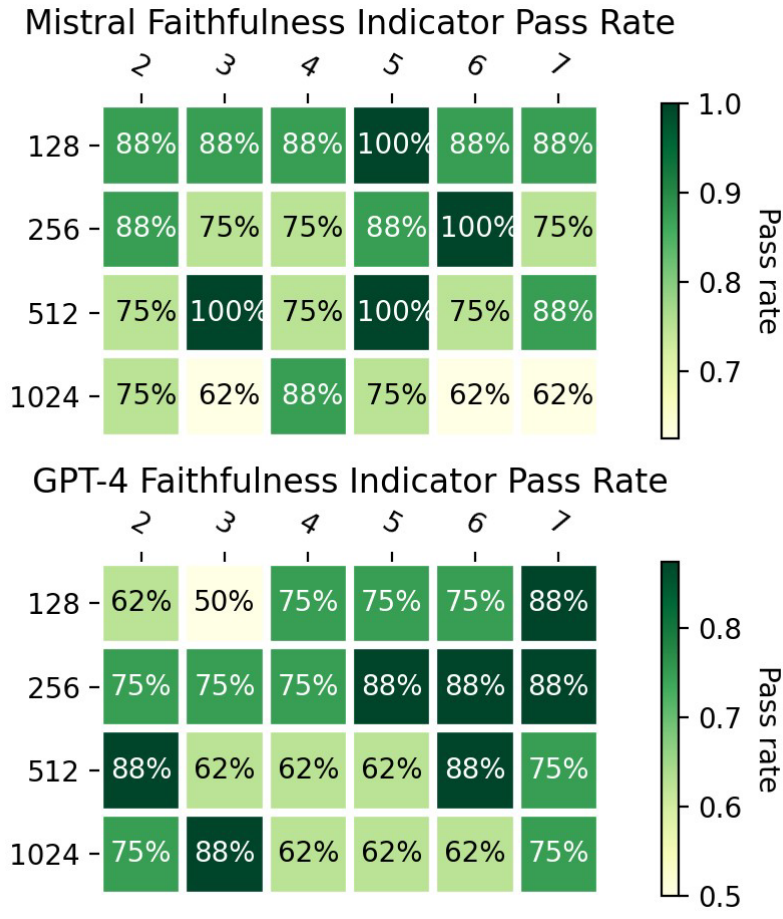


Figure 1-7. Mistral-7b and GPT-4 Faithfulness Passing Rates by Chunk Size and Number of Documents Retrieved

2. Response Comparison

In this section, we analyze the differences between specific responses from both GPT-4 and Mistral-7b under the same chunk size and document retrieved settings. We include three examples: one in which both LLMs passed the metrics, one in which they both failed, and one in which GPT-4 passed Relevancy but Mistral-7b did not.

A. Example 1

For Prompt 7, “What work in Taiwan is being funded?”, GPT-4 passed both Faithfulness and Relevancy 96% of the time, and Mistral-7b passed Faithfulness every time and Relevancy 79% of the time. The following responses passed both metrics using a chunk size of 256 and three retrieved documents.

1. GPT-4

The work in Taiwan being funded includes strengthening the partnership with Taiwan, reform and integration of Taiwan's reserve military forces, the use of defense articles and services transferred from the United States to Taiwan, the integration of Taiwan's military forces with relevant civilian agencies, including the All-Out Defense Mobilization Agency, enhancing Taiwan's ability to participate in bilateral and multilateral military exercises, improving the defensive cyber capabilities and practices of the Ministry of National Defense of Taiwan, and other matters considered relevant by the Secretary of Defense. Additionally, there is an establishment of a comprehensive training, advising, and institutional capacity-building program for the military forces of Taiwan.

2. Mistral-7b

The United States government is funding various activities in Taiwan under the Taiwan Relations Act and other applicable statutory authorities. These activities include but are not limited to strengthening the United States partnership with Taiwan, reform and integration of the reserve military forces of Taiwan, use of defense articles and services transferred from the United States to Taiwan, integration of the military forces of Taiwan with relevant civilian agencies, ability of Taiwan to participate in bilateral and multilateral military exercises, defensive cyber capabilities and practices of the Ministry of National Defense of Taiwan, and any other matter the Secretary of Defense considers relevant.

3. Analysis

These responses included the actions that the DoD can take under the Taiwan Relations Act and ways that the two countries' militaries can cooperate. These specific actions, such as joint military activities and improving cyber capabilities, are cited by both GPT-4 and Mistral-7b and used in a way that directly answers the question. This question is broader than some of the other questions used for this exercise, as there are more relevant documents that can be retrieved to help with the answer and the answer does not need to be very specific to be relevant.

B. Example 2

For Prompt 8, "Which study involving recruiting and retention is the highest priority for a federally funded research and development center (FFRDC)? Explain why," Mistral-7b outperformed GPT-4 on the Faithfulness metric, passing 83% of the time compared to GPT-4's 42%. However, GPT-4 did much better on Relevancy, passing 38% of the time compared to Mistral-7b's 13%. The following two responses failed both metrics. The LLMs made these responses using a chunk size of 512 tokens and six retrieved documents.

1. GPT-4

The highest priority study involving recruiting and retention for a federally funded research and development center (FFRDC) is not explicitly ranked in the provided context. The context information outlines various programs, evaluations, and competitions, but it does not specify a prioritization of studies specifically related to recruiting and retention for FFRDCs. Therefore, based on the information given, it is not possible to determine which study is the highest priority in this area.

2. Mistral-7b

Based on the information provided in the query, it appears that the FFRDC is conducting a study related to recruiting and retention. However, without further context or information, it is not possible to determine which specific study within this broader category is the highest priority for the FFRDC. Additionally, it is important to note that the priorities of an FFRDC can change based on various factors, including changes in federal funding levels, shifts in national security needs and priorities, and advances in scientific knowledge and technology. Therefore, it is essential that any determination of a priority for an FFRDC be made with a clear understanding of the current context and the potential for future changes. In summary, based on the information provided in the query, it is not possible to determine which specific study within the broader category of studies related to recruiting and retention is the highest priority for the FFRDC. Any determination of a priority for an FFRDC must be made with a clear understanding of the current context and the potential for future changes.

3. Analysis

Neither of these responses directly answer which study might be the highest priority for an FFRDC even though the context information includes examples of possible FFRDC work, such as studies of recruiting strategies for mining schools that might compete for federal contracts or personnel management research at the National Academies. These documents are not specifically referenced in these responses, and no determination is made over which study or studies is highest priority, which likely explains why these responses failed both the Faithfulness and Relevancy metrics. This question may have been challenging for both GPT-4 and Mistral-7b, as it asks them to determine something that is not specifically stated in the NDAA as opposed to other questions about specific content. Additionally, even though relevant information was included, the LLMs retrieved some unrelated documents, such as a section on retention boards for the Space Force.

C. Example 3

For Prompt 6, “What studies involving recruiting and retention are in the NDAA?”, both GPT-4 and Mistral-7b passed Faithfulness over 90% of the time. However, Mistral-7b passed Relevancy only 33% of the time, whereas GPT-4 passed it 83% of the time. In the following responses, both LLMs passed Faithfulness, but only GPT-4 passed Relevancy. The LLMs made these responses using a chunk size of 256 with five retrieved documents.

1. GPT-4

The studies involving recruiting and retention in the NDAA include:

1. A report that includes the name, military service, and sport of each military service graduate released or deferred from active service to participate in professional sports, along with a description of their sports career progress, a summary of marketing strategy and recruiting related activities conducted, a description of assessments conducted by the military services to determine the recruiting value associated with approved releases from active duty, and the current status of each participant.
2. An evaluation of a program that involves identifying foreign students recommended for admission under the Senior Reserve Officers Training Corps program, coordinating with partner countries to evaluate interest and promote awareness of the program, establishing a mechanism for tracking an alumni network of foreign students who participate in the program, and working with participant institutions and partner countries to identify academic institutions and programs of interest to participating countries.
3. A report on the program mentioned above, which includes a narrative summary of activities conducted as part of the program during the preceding fiscal year, an overview of participant Senior Reserve Officers Training Corps programs, individuals, and countries, a description of the areas of study entered into by the students participating in

the program, and a description of opportunities and potential timelines for future program expansion.

2. Mistral-7b

The National Defense Authorization Act (NDAA) for Fiscal Year 2023 includes provisions related to recruiting and retention of military personnel. Specifically, under Section 514, the Secretary of Defense is required to establish a program to provide educational opportunities to foreign students who are military service graduates and have been released or deferred from active service in order to participate in professional sports. The report required under this section includes information on the sports career progress of each participant, such as drafted, signed, re-leased, or returned to military service.

3. Analysis

Both of these responses refer to studies on military service graduates released from or deferring service to participate in sports and on admission of foreign students into senior ROTC. This likely explains why they pass the Faithfulness metric. However, while GPT-4 correctly splits these into separate studies, Mistral-7b merges them into one study involving foreign students participating in sports, which is not correct and not relevant to what the question is asking for. This is likely why GPT-4 passes the Relevancy test while Mistral-7b fails.

D. Similarity Scores

To guide our understanding of the document retrieval process, we also reviewed the similarity scores associated with document retrieval by chunk size and used the Euclidean (L2) distance as the distance function. Thus, the lowest similarity score indicates that the query was close to another object embedded in the vector library by L2 distance. The following experiments use different prompts to provide context to the similarity scores. Table 2-1 details our use of NDAA Sec. 2203, which we stored as a single chunk in the vector library for chunk size 1,024:

SEC. 2203. AUTHORIZATION OF APPROPRIATIONS, NAVY. (a) AUTHORIZATION OF APPROPRIATIONS.--Funds are hereby authorized to be appropriated for fiscal years beginning after September 30, 2023, for military construction, land acquisition, and military family housing functions of the Department of the Navy, as specified in the funding table in section 4601. (b) LIMITATION ON TOTAL COST OF CONSTRUCTION PROJECTS.--Notwithstanding the cost variations authorized by section 2853 of title 10, United States Code, and any other cost variation authorized by law, the total cost of all projects carried out under section 2201 of this Act may not exceed the total amount authorized to be appropriated under subsection (a), as specified in the funding table in section 4601.

Table 2-1. Similarity Scores for Select Chunks of Text

Prompt text	Similarity score for chunk size 1,024
Exact text of NDAA Sec. 2203	0
SEC. 2203. AUTHORIZATION OF APPROPRIATIONS, NAVY. (a) AUTHORIZATION OF APPROPRIATIONS.—Funds are hereby authorized to be appropriated for fiscal years beginning after September 30, 2023, for military construction, land acquisition, and military family housing functions of the Department of the Navy, as specified in the funding table in section 4601.	0.256
Section 2203 states that Department of the Navy funding is authorized to be appropriated for fiscal years beginning after September 30, 2023 for military construction, land acquisition, and military family housing, as specified in the funding table in Section 4601. There is also a limitation on total cost of construction projects.	0.329
Will there be a limitation on total cost of construction projects in the 2023 authorization of appropriations for the Navy?	0.386

In this context, 0 can be interpreted as an exact chunk, scores in the range from 0.26–0.33 represent a partial copy or synopsis of one chunk, and a higher score represents a question directly referencing a single chunk. In all cases, the top document retrieved was the exact chunk itself. For our RAG experiment, we did not take any prompts directly from the NDAA, so the similarity scores will be higher.

Table 2-2. Minimum Similarity Score by Prompt and Document Chunk Size

Prompt	128	256	512	1,024
1. Tell me about artificial intelligence in the Coast Guard.	0.746	0.811	0.857	0.857
	1.007	0.973	1.016	1.016

Prompt	128	256	512	1,024
2. What are key concerns for the Space Force?	0.963	0.950	0.953	0.953
3. What are some concerns for the Space Force that outside government contractors may be needed to address?	0.646	0.838	0.881	0.918
4. What is said about amphibious vehicles?	0.830	0.867	0.885	0.885
5. What sections mention the military health system?	0.997	1.139	1.200	1.186
6. What studies involving recruiting and retention are in the NDAA?				
7. What work in Taiwan is being funded?	0.895	0.889	1.055	1.076
8. Which study involving recruiting and retention is the highest priority for an FFRDC? Explain why.	0.853	1.089	1.121	1.149

Across all questions, as shown in Table 2-2 above, the similarity score of the highest-ranked document in the retrieval system increases as the chunk size increases. The mean of the minimum similarity score across all questions was 0.867 for the lowest chunk size and 1.005 for the highest chunk size. This 16% increase demonstrates that the highest-ranked document retrieved was typically judged to be less relevant by the retrieval system for larger chunk sizes. Additionally, similarity scores largely did not differ between the two largest chunk sizes.

1. What is said about amphibious vehicles?

While this question had the best similarity score for any question-chunk size pair for the 128-token chunk size and had low similarity scores in general, the metrics show that the GPT-4 model particularly struggled to answer it. GPT-4's responses had the lowest pass rate from the Faithfulness metric and the second lowest pass rate from the Relevancy metric. The retrieved NDAA text frequently pertained to amphibious landing docks or other amphibious ships with unclear relationships to the question. An amphibious vehicle is defined as a vehicle that can operate both on land and sea, and not all amphibious warships are amphibious vehicles. This suggests that the distinction between amphibious vehicles and warships is unclear within the embedding model, contributing to GPT-4's poor performance.

2. Tell me about artificial intelligence in the Coast Guard.

Overall, the minimum similarity scores by chunk size were the lowest for this question. However, the documents that were retrieved for this question were largely unhelpful. The top two documents retrieved were portions of Section 222 regarding plans and strategies for artificial intelligence in the DoD and portions of Section 332 regarding the use of artificial intelligence for shipyard optimization. The Coast Guard is not part of the Department of Defense, so these documents referencing Section 222 do not pertain to the question despite the low similarity score. Likewise, Section 332 pertains to Navy policy, which is largely irrelevant to the Coast Guard. This indicates that the embedding model could be fine-tuned on specific DoD data that clarifies the distinction between the Coast Guard, the DoD, and the Navy.

3. What sections mention the military health system?

For this question, the similarity scores were lower than most of the other questions for all chunk sizes except for the 128-token one and matched relevant documents that led the LLMs to provide a good answer. Note that this prompt requests the specific sections (essentially word for word) rather than requiring a lot of inference or reasoning from the LLMs.

For the other cases, the document retriever could not find relevant documents according to the embedding model, noted by the higher similarity scores. This could help to explain inadequacies in the LLMs' response.

E. Limitations

The document ingestion process is a clear candidate for changes to improve query results. We varied only the chunk size and the number of documents retrieved. Other potentially useful parameters include the way that the document was split and the chunk overlap length. The current method of splitting the document involves only a carriage return. This is often used in RAG to keep paragraphs together as most documents are organized that way. However, the NDAA is organized by sections and subsections, so it may be more useful to look for characters such as parentheses or other NDAA-specific indicators. Additional cleaning of the NDAA text such as removal of page numbers or other text that does not add meaning, could also improve results.

For embedding, we used HuggingFace's all-MiniLM-L6-v2 sentence and short paragraph transformer model. This model is optimized for speed and is intended to be a general-purpose model. There are other models that are intended for Q&A, such as multi-qa-distilbert-cos-1 or multi-qa-MiniLM-L6-cos-v1 that could be of more utility for this specific use case. Additionally, all-MiniLM-L6-v2 has an input text limit of 256 words at a time, while other models have larger input text limits that may be better for our use case, such as multi-qa-MiniLM-L6-cos-v1 with its 512-word chunk limit. Further, as these models have all been trained on a general-purpose data set, fine tuning them on a defense-oriented or NDAA-specific dataset may increase performance. Lastly, we used an uncased model, meaning that "coast guard" has the same semantic meaning as

“Coast Guard.” Although we do not know whether a cased model would increase performance, it is an avenue to explore.

The retrieval method could also be varied. We used the FAISS library as a vector retriever, which uses L2 distance as the similarity search method. Further, we used the top-k document retrieval method versus a similarity score threshold. One could apply re-ranking techniques or generate an ideal response or document to improve document retrieval. This could strongly impact performance and should be considered in future work.

Finally, the method of evaluation used for this analysis can be varied as well. There are packages beyond LlamaIndex that use LLMs to evaluate RAG responses, such as those offered by Galileo or Ragas. Another way to evaluate responses and retrieval is to create a set of correct answers and then check the RAG responses against the correct set. The drawback to this approach is that the correct answer set is subjectively determined by the human evaluator.

F. Conclusions

When evaluating the performance of RAG systems or of LLM systems in general, it is important to evaluate both the LLM response and the retrieval system. Overall, the responses performed better by relevancy when using chunk sizes between 128 and 256 tokens. Likewise, the documents retrieved had improved similarity scores on average when using smaller chunk sizes. We found no clear pattern for chunk size or number of documents retrieved that appeared to impact the results of the Faithfulness metric.

By and large, the results of the Faithfulness and Relevancy metrics indicate that the quality of the responses also seemed to be driven by the nature of the question. Mistral-7b and GPT-4 had higher overall pass rates for questions with answers directly mentioned in the text, such as, “What work in Taiwan is being funded?” or “What sections mention the military health system?” However, Mistral-7b struggled with the Relevancy metric, failing over half the time on average across all eight prompts. GPT-4 was more successful, failing closer to 20% of the time. In terms of the Faithfulness metric, both models did well, although GPT-4 notably struggled more with certain questions, such as the amphibious vehicles question. Also, the metrics used for this analysis could be difficult to interpret for questions lacking a clear answer in the NDAA, so it is possible that other methods of evaluating a RAG system could have different results.

Lastly, the models worked at very different speeds. The Mistral-7b model averaged more than 15 times longer to generate a response for the prompts with the most context compared to those with the least context. GPT-4’s difference in response time was much less drastic, taking only about 3 times longer. However, we ran Mistral-7b entirely on CPU inference, so the difference in speed could potentially be greatly reduced with a GPU-based system.

Although RAG systems show potential for allowing staff to query the NDAA in natural language, they will need to be closely monitored for reliability across different metrics before

deployment. Additionally, more research will be needed on the impact of GPU inference and varied document retrieval and prompting techniques.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188		
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YY) 00-05-24		2. REPORT TYPE Final		3. DATES COVERED (From – To)	
4. TITLE AND SUBTITLE Talk to the NDAA: Evaluating Retrieval Augmented Generation (RAG) on Congressional Legislation using Large Language Models (LLM)			5a. CONTRACT NUMBER HQ0034-19-D-0001		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBERS		
6. AUTHOR(S) William M. Fisher, Astrid I. Berge			5d. PROJECT NUMBER C5240		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES Institute for Defense Analyses 730 East Glebe Road Alexandria, VA 22305			8. PERFORMING ORGANIZATION REPORT NUMBER IDA Product 3001199		
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 730 East Glebe Road, Alexandria, VA 22305			10. SPONSOR'S / MONITOR'S ACRONYM IDA		
			11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A. Approved for public release: distribution is unlimited.					
13. SUPPLEMENTARY NOTES Project Leader: Kevin Garrison					
14. ABSTRACT We explore the use of Large Language Models (LLM) and Retrieval Augmented Generation methods to summarize the draft National Defense Authorization Act for Fiscal Year 2024.					
15. SUBJECT TERMS Large language model, retrieval augmented generation, hallucination, prompting, NDAA, LLM evaluation					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unlimited	18. NUMBER OF PAGES 21	19a. NAME OF RESPONSIBLE PERSON Institute for Defense Analyses
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include Area Code)

