Evaluating Information Relevance in Retrieval-Augmented Generation

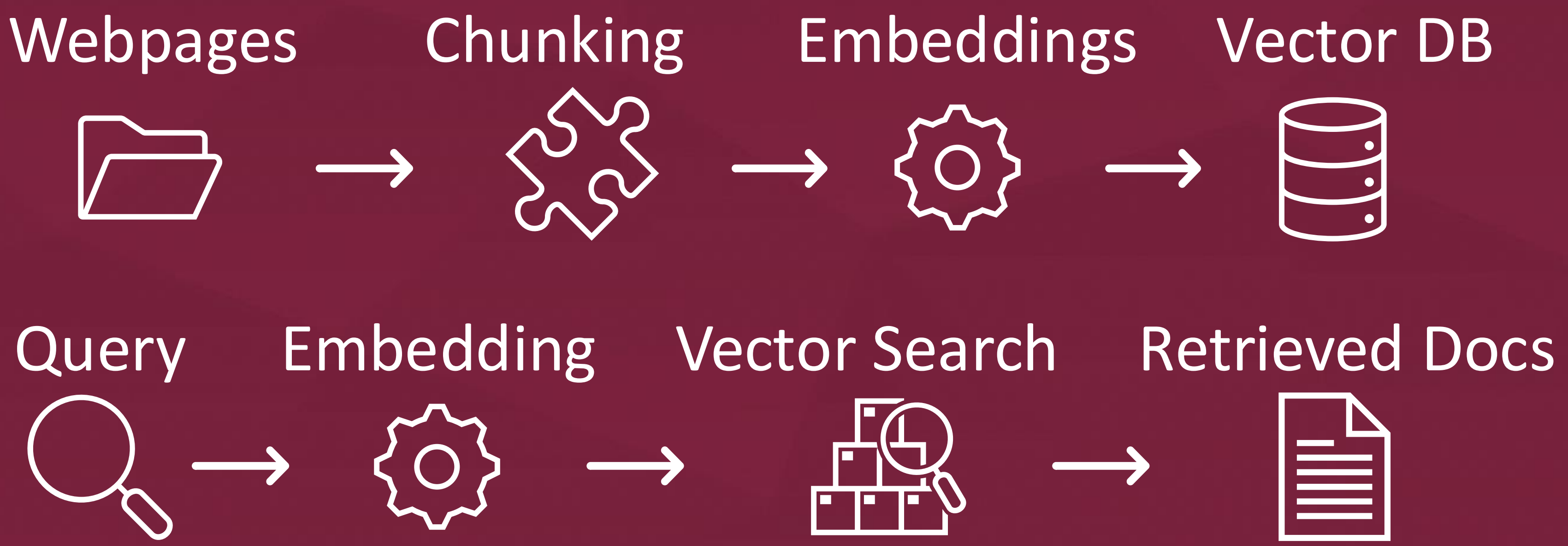
Fangyun Yang, Anton Picetti
Marius Bobinas, Mahbubur Rahman

BACKGROUND

- Large Language Models rely on static training data
- Prohibitively expensive and impractical retraining
- Risk of hallucinations and irrelevant outputs

SETUP

- Incorporating external knowledge sources
- Embedding documents and user queries (all-MiniLM-L6-v2)
- Semantic vector search for retrieval



Webpages → Chunking → Embeddings → Vector DB

Query → Embedding → Vector Search → Retrieved Docs

Project question

Can RAG performance be improved using Chain-of-Thought reasoning, and further enhanced with a self-correcting state-graph mechanism?

RAG

- Retrieval-Augmented Generation
- Combines a retrieval system and a large language model
- Indirectly answers
- Up-to-date information->knowledge base
- LLM
- Grounded answers

Offline stage

- Documents collected,and converted into vector representaions
- Stored in a vector database

Online stage

- Ask question
- Retrieve the the relevant documents
- Pass them to the language model
- Generate the final answer

Document Processing

Embeddings

Vector Database

Retriever

RAG

LLM

Evaluation

Challenges with RAG

Does the response answer the question?

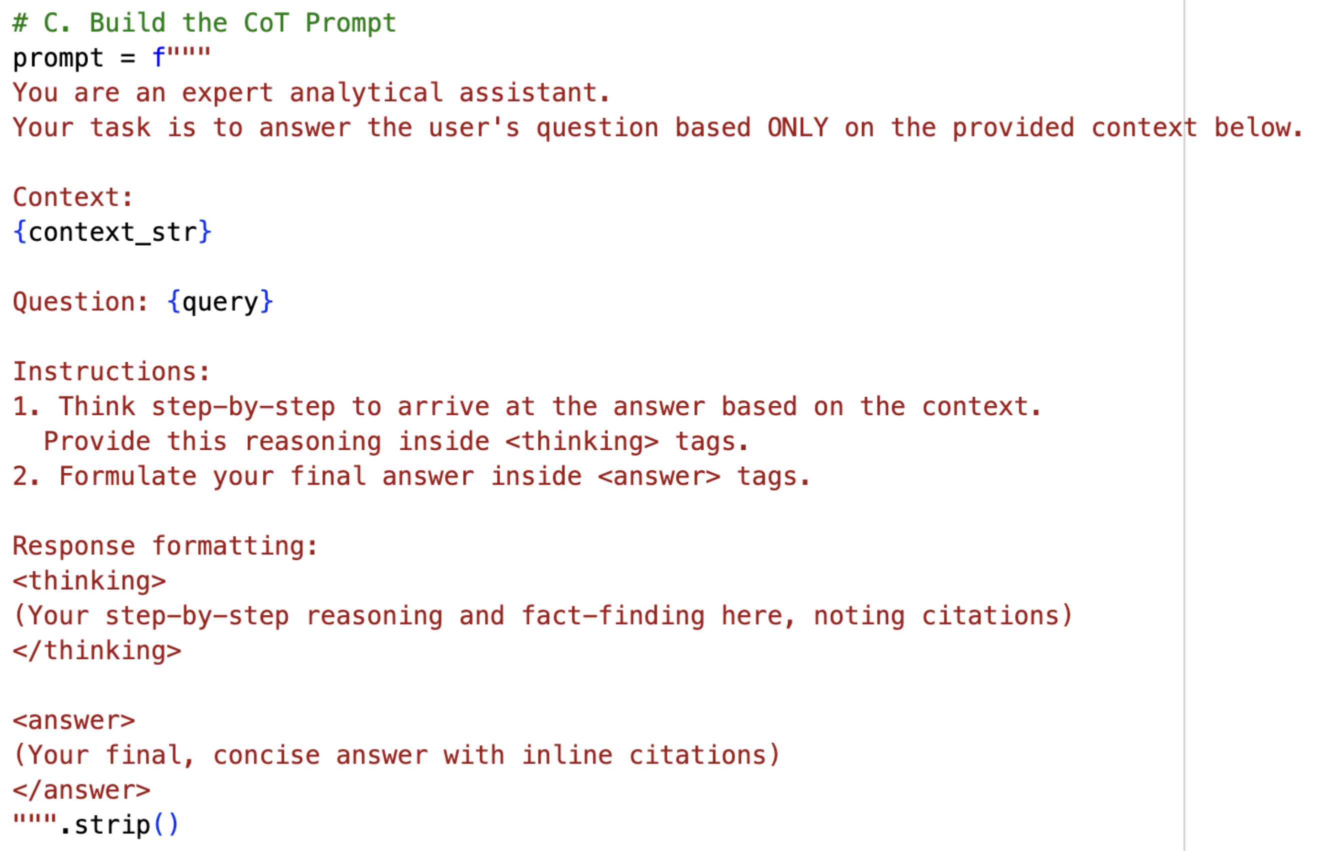
Are the claims in the response supported by the retrieved context?

Are irrelevant claims included in the response?

Is the response aligned with the ground truth?

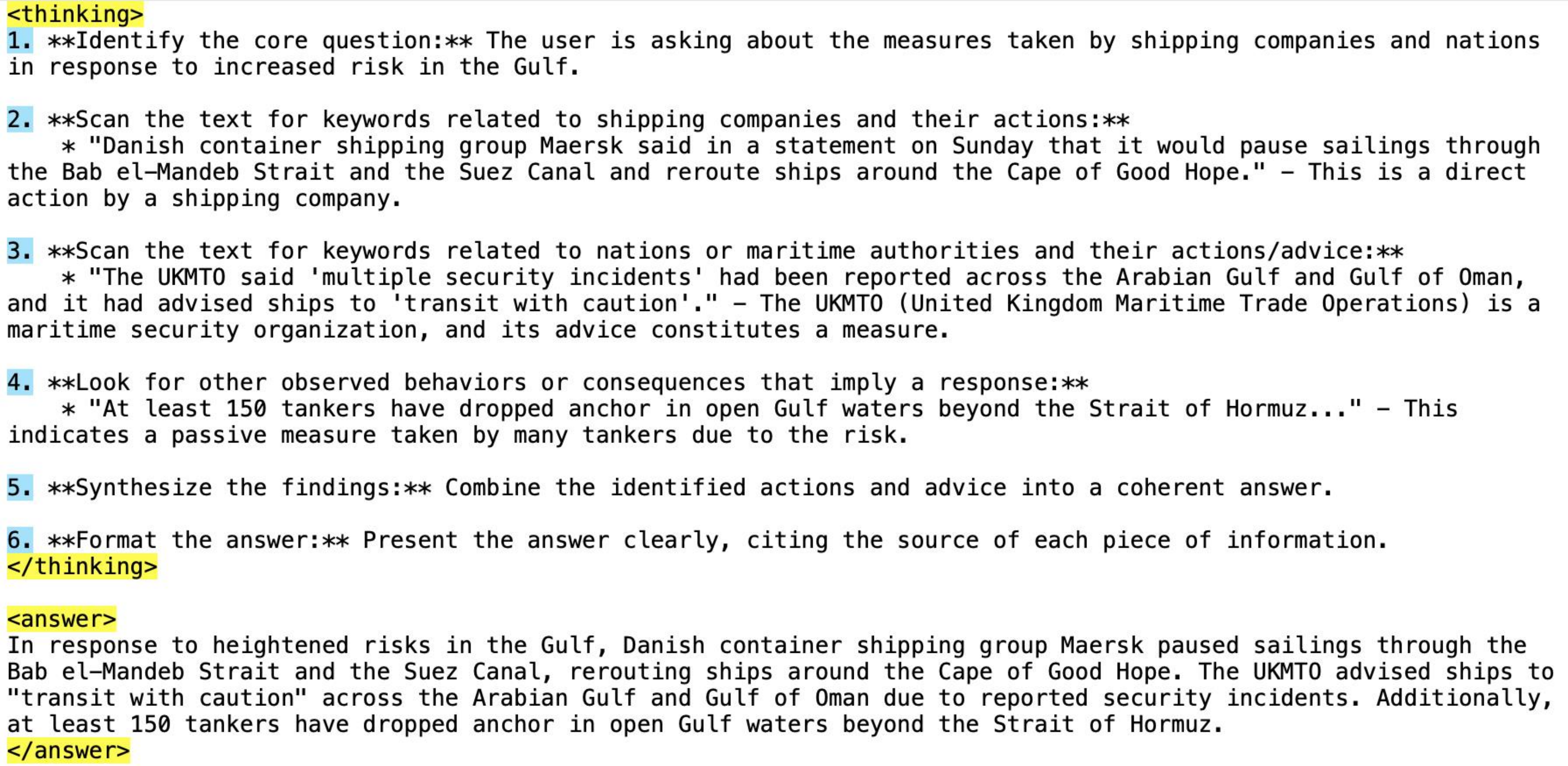
Promting for CoT

- Encourage the model to reason step by step:



```
# C. Build the CoT Prompt
prompt = f"""
You are an expert analytical assistant.
Your task is to answer the user's question based ONLY on the provided context below.

Context:
{context_str}

Question: {query}

Instructions:
1. Think step-by-step to arrive at the answer based on the context.
  Provide this reasoning inside <thinking> tags.
2. Formulate your final answer inside <answer> tags.

Response formatting:
<thinking>
(Your step-by-step reasoning and fact-finding here, noting citations)
</thinking>

<answer>
(Your final, concise answer with inline citations)
</answer>
""".strip()
```

Promting for CoT

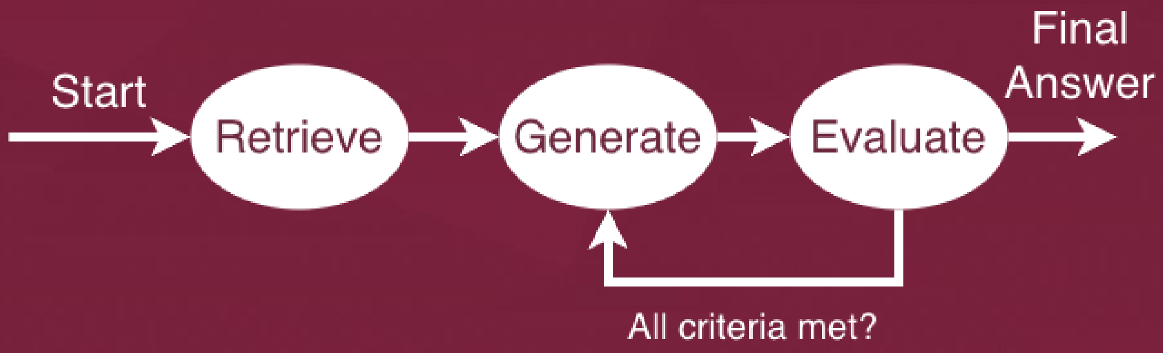


```
<thinking>
1. **Identify the core question:** The user is asking about the measures taken by shipping companies and nations
in response to increased risk in the Gulf.

2. **Scan the text for keywords related to shipping companies and their actions:**
    * "Danish container shipping group Maersk said in a statement on Sunday that it would pause sailings through
the Bab el-Mandeb Strait and the Suez Canal and reroute ships around the Cape of Good Hope." - This is a direct
action by a shipping company.

3. **Scan the text for keywords related to nations or maritime authorities and their actions/advice:**
    * "The UKMTO said 'multiple security incidents' had been reported across the Arabian Gulf and Gulf of Oman,
and it had advised ships to 'transit with caution'." - The UKMTO (United Kingdom Maritime Trade Operations) is a
maritime security organization, and its advice constitutes a measure.

4. **Look for other observed behaviors or consequences that imply a response:**
    * "At least 150 tankers have dropped anchor in open Gulf waters beyond the Strait of Hormuz..." - This
indicates a passive measure taken by many tankers due to the risk.

5. **Synthesize the findings:** Combine the identified actions and advice into a coherent answer.

6. **Format the answer:** Present the answer clearly, citing the source of each piece of information.
</thinking>

<answer>
In response to heightened risks in the Gulf, Danish container shipping group Maersk paused sailings through the
Bab el-Mandeb Strait and the Suez Canal, rerouting ships around the Cape of Good Hope. The UKMTO advised ships to
"transit with caution" across the Arabian Gulf and Gulf of Oman due to reported security incidents. Additionally,
at least 150 tankers have dropped anchor in open Gulf waters beyond the Strait of Hormuz.
</answer>
```

Custom CoT

Criteria:
- Is the answer **faithful**?
- Is the answer **relevant**?
- Is the answer **correct**?

- A simple LangGraph architecture
- Prompting based CoT relies on a single model response
- Iterative self-correction mechanism
- State graph where each new state results in a new LLM-call

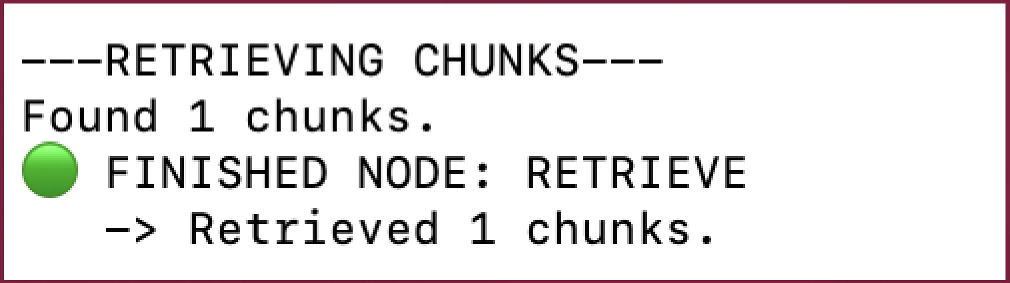


Custom CoT – Example use case

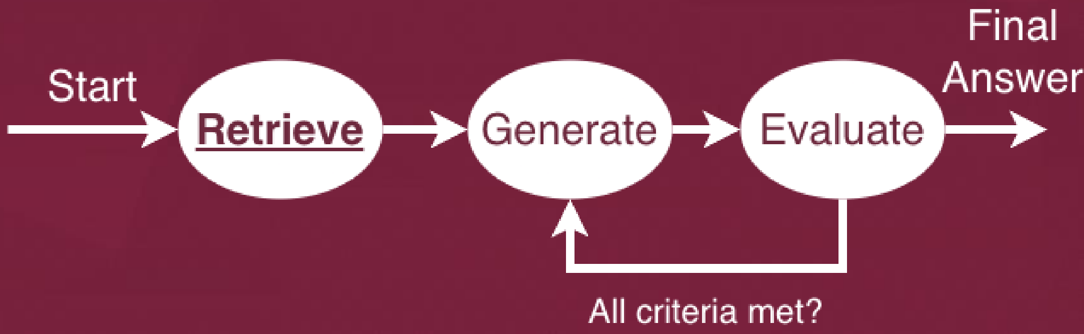
- Article: Oil prices rise after ships attacked near Strait of Hormuz
- Question: What measures did shipping companies **and** nations take in response to the heightened risk in the Gulf?

Custom CoT – Example use case

Retrieving data from vector database

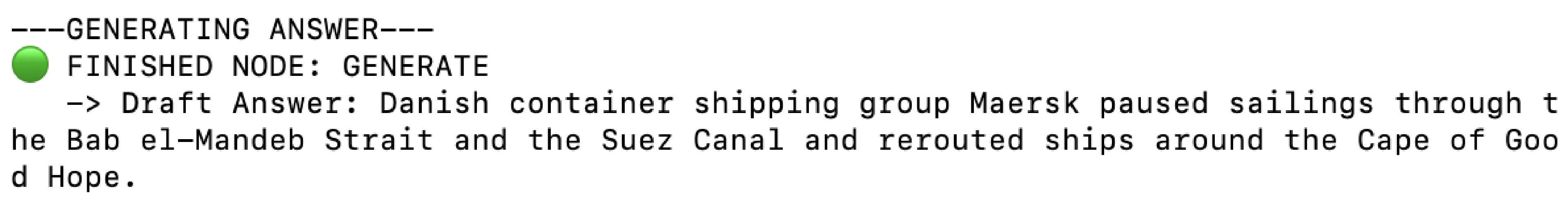


```
---RETRIEVING CHUNKS---
Found 1 chunks.
🟢 FINISHED NODE: RETRIEVE
   -> Retrieved 1 chunks.
```

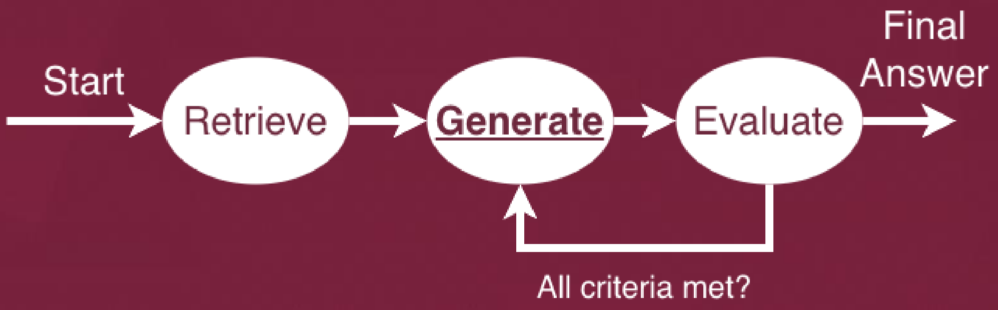


Custom CoT – Example use case

Generate draft

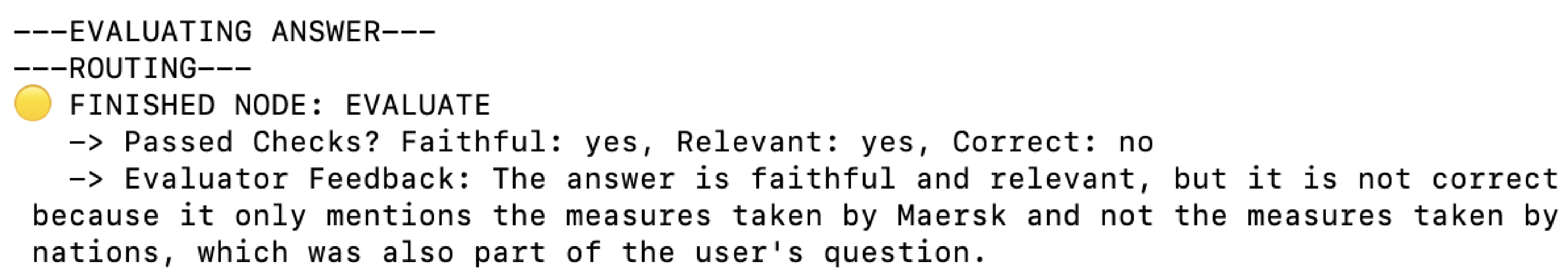


```
---GENERATING ANSWER---
🟢 FINISHED NODE: GENERATE
   -> Draft Answer: Danish container shipping group Maersk paused sailings through t
he Bab el-Mandeb Strait and the Suez Canal and rerouted ships around the Cape of Goo
d Hope.
```

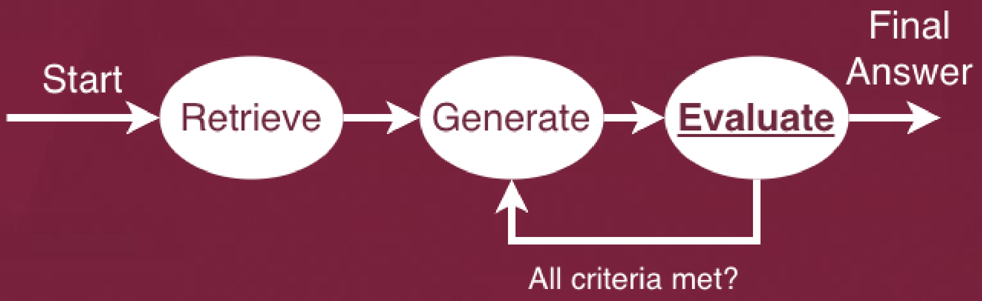


Custom CoT – Example use case

Evaluate draft

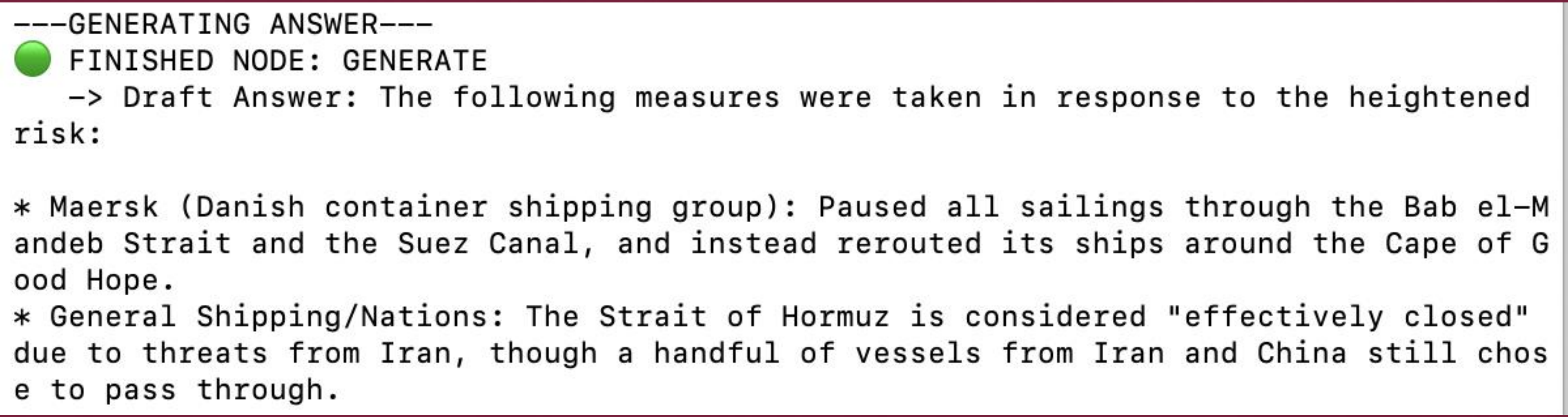


```
---EVALUATING ANSWER---
---ROUTING---
🟡 FINISHED NODE: EVALUATE
   -> Passed Checks? Faithful: yes, Relevant: yes, Correct: no
   -> Evaluator Feedback: The answer is faithful and relevant, but it is not correct
 because it only mentions the measures taken by Maersk and not the measures taken by
 nations, which was also part of the user's question.
```

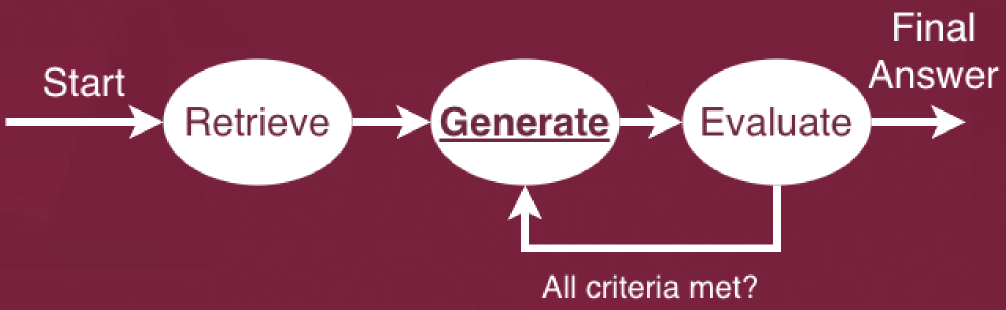


Custom CoT – Example use case

Generate second draft

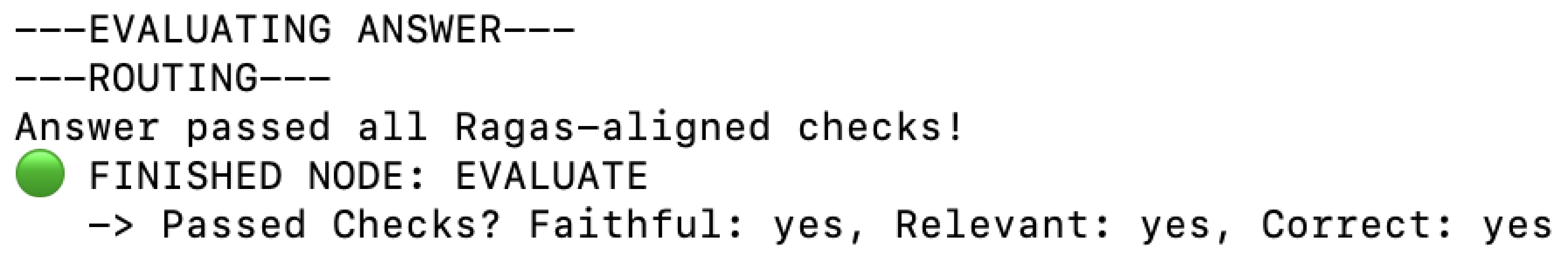


```
---GENERATING ANSWER---
🟢 FINISHED NODE: GENERATE
   -> Draft Answer: The following measures were taken in response to the heightened
risk:

* Maersk (Danish container shipping group): Paused all sailings through the Bab el-M
andeb Strait and the Suez Canal, and instead rerouted its ships around the Cape of G
ood Hope.
* General Shipping/Nations: The Strait of Hormuz is considered "effectively closed"
due to threats from Iran, though a handful of vessels from Iran and China still chos
e to pass through.
```

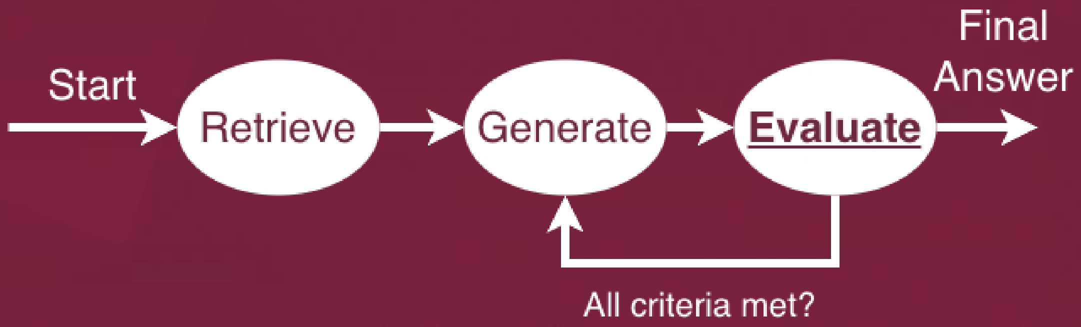


Custom CoT – Example use case

Evaluate second draft

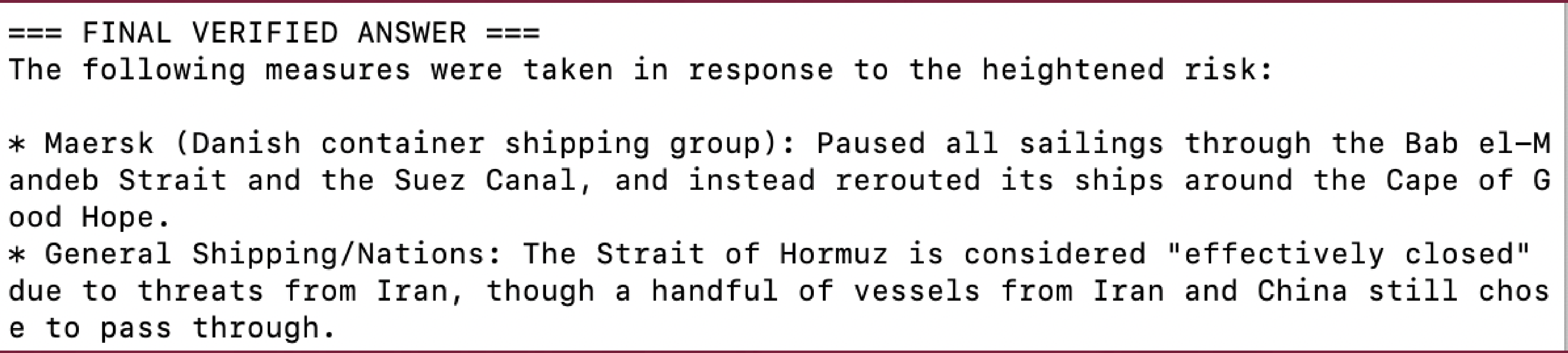


```
---EVALUATING ANSWER---
---ROUTING---
Answer passed all Ragas-aligned checks!
🟢 FINISHED NODE: EVALUATE
   -> Passed Checks? Faithful: yes, Relevant: yes, Correct: yes
```

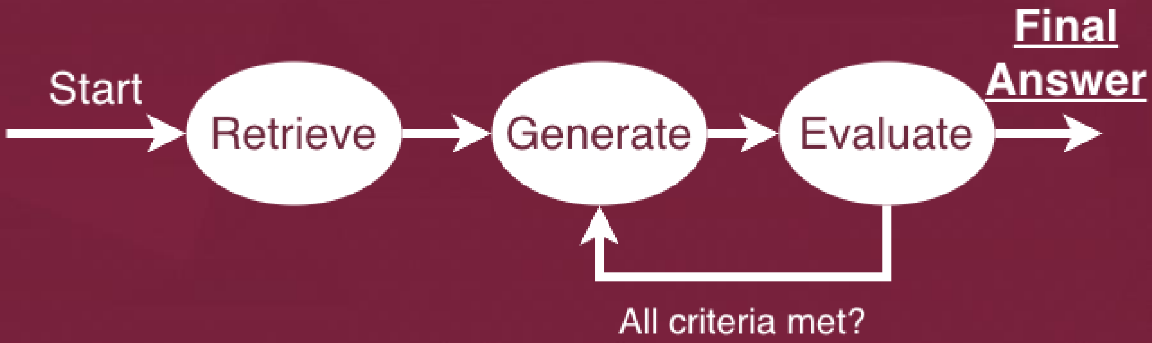


Custom CoT – Example use case

Final answer



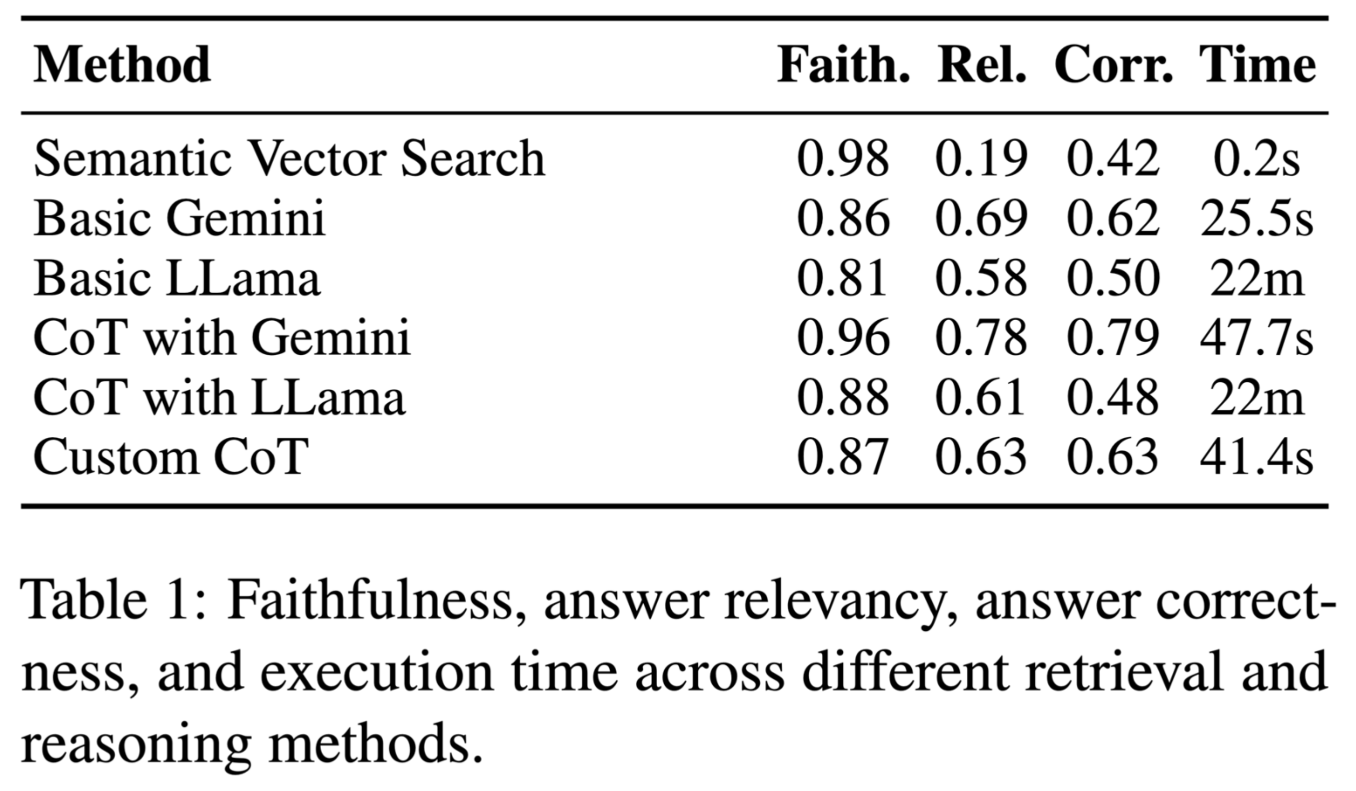
```
=== FINAL VERIFIED ANSWER ===
The following measures were taken in response to the heightened risk:

* Maersk (Danish container shipping group): Paused all sailings through the Bab el-M
andeb Strait and the Suez Canal, and instead rerouted its ships around the Cape of G
ood Hope.
* General Shipping/Nations: The Strait of Hormuz is considered "effectively closed"
due to threats from Iran, though a handful of vessels from Iran and China still chos
e to pass through.
```



Evaluation with RAGAS

- Faithfulness: “Is the answer supported by the retrieved context?”
- Answer Relevancy: “Does the answer actually address the question?”
- Answer Correctness: “Is the answer factually correct compared to reference?”

Results & Analysis



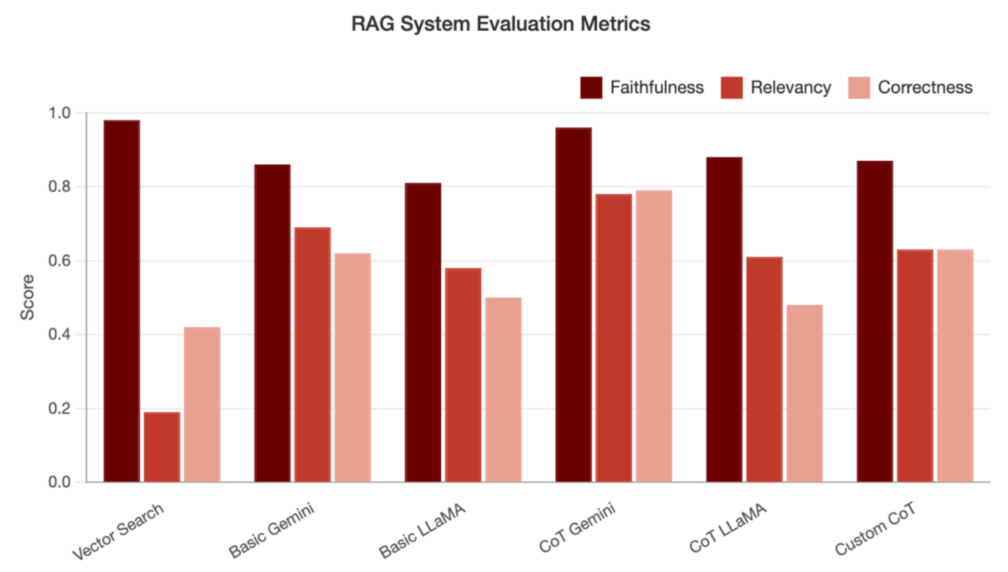
Method	Faith.	Rel.	Corr.	Time
Semantic Vector Search	0.98	0.19	0.42	0.2s
Basic Gemini	0.86	0.69	0.62	25.5s
Basic LLama	0.81	0.58	0.50	22m
CoT with Gemini	0.96	0.78	0.79	47.7s
CoT with LLama	0.88	0.61	0.48	22m
Custom CoT	0.87	0.63	0.63	41.4s

Table 1: Faithfulness, answer relevancy, answer correctness, and execution time across different retrieval and reasoning methods.

Results & Analysis

Vector search:

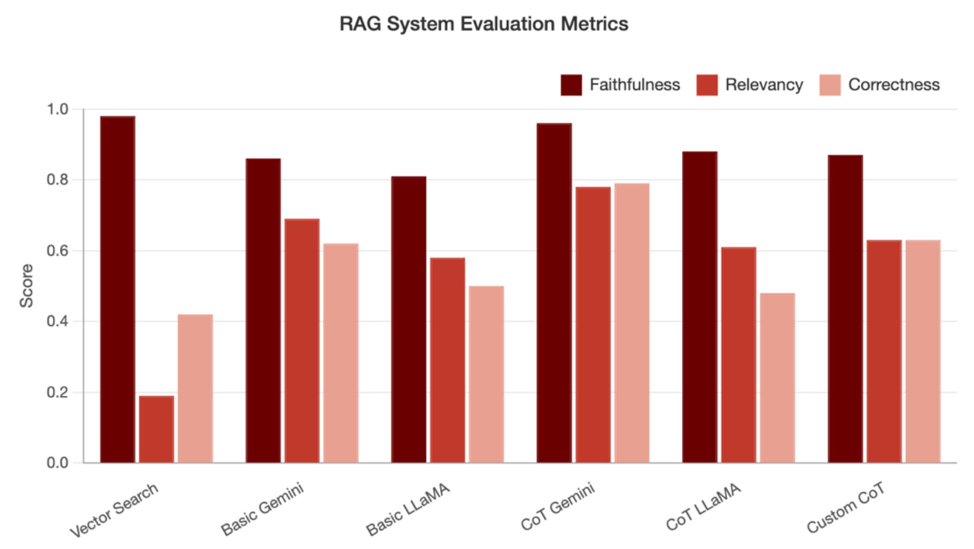
- Retrieved content is accurate
- But not answering the question properly



Results & Analysis

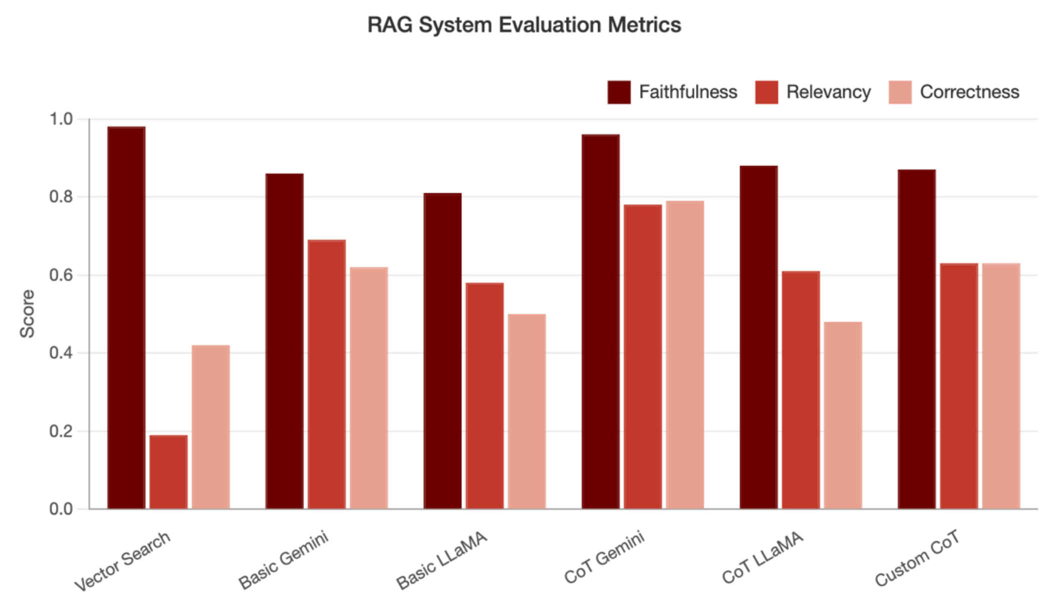
RAG with Gemini
- Good balance of metrics and Fast

RAG with LLAMA
- Slower but more relevant


RAG System Evaluation Metrics

Results & Analysis

- The best performance comes from CoT with Gemini
- CoT with LLaMA performs slightly worse
- Reasoning ability depends heavily on model capability



Potential Improvements

- Larger datasets: testing across more diverse news articles & queries
- Larger local models
- Prompt optimization: prompt tuning for specific domains

Conclusion

- RAG improves grounding
- CoT improves reasoning with strong models
- Performance might drop with weaker models

Thanks.
Questions?