This table displays various RAG LLM models and their parameters, along with two baseline models for comparison.

Model Number	Model Name	Parameters	Documentation Sources	Prompt Techniques	Re-ranking	Dual-LLM Design	Embedding Model	LLM Model
1	RAG Model 1	chunk_size = 1000 chunk_overlap = 400 k = 4 fetch_k = 0 mmr: False	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1
2	RAG Model 2	chunk_size = 1000 chunk_overlap = 400 k = 6 fetch_k = 0 mmr: True	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1
3	RAG Model 3	chunk_size = 1500 chunk_overlap = 600 k = 6 fetch_k = 0 mmr: True	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1
4	RAG Model 4	chunk_size = 2000 chunk_overlap = 900 k = 6 fetch_k = 0 mmr: True	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1
5	RAG Model 5	chunk_size = 1500 chunk_overlap = 600 k = 8 fetch_k = 0 mmr: True	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1

Model Number	Model Name	Parameters	Documentation Sources	Prompt Techniques	Re-ranking	Dual-LLM Design	Embedding Model	LLM Model
6	RAG Model 6	chunk_size = 1500 chunk_overlap = 400 k = 8 fetch_k = 20 mmr: True	Chameleon Blogs (Selected) ReadtheDocs (Index Pages Only) FAQs Forum Posts	Simple Prompt	None	None	BAAI/bge-large-en	Llama3.1
7	RAG Model 7	chunk_size = 1000 chunk_overlap = 400 k = 6 fetch_k = 20 mmr: True	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt	None	None	BAAI/bge-large-en	Llama3.1
8	RAG Model 8	chunk_size = 1000 chunk_overlap = 150 k = 10 fetch_k = 20 mmr: True	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt	None	None	BAAI/bge-large-en	Llama3.1
9	RAG Model 9	chunk_size = 2000 chunk_overlap = 150 k = 5 fetch_k = 10 mmr: True	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt	None	None	BAAI/bge-large-en	Llama3.1
10	RAG Model 10	chunk_size = 512 chunk_overlap = 100 k = 7 fetch_k = 28 mmr: True	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt	None	None	BAAI/bge-large-en	Llama3.1
11	RAG Model 11	chunk_size = 1024 chunk_overlap = 150	Chameleon Blogs (Full) ReadtheDocs (All Pages)	Enhanced Prompt	BAAI/bge-reranker-large	None	BAAI/bge-base-en-v1.5	Llama3.1

Model Number	Model Name	Parameters	Documentation Sources	Prompt Techniques	Re-ranking	Dual-LLM Design	Embedding Model	LLM Model
		k = 5 fetch_k = 20 mmr: False	FAQs Forum Posts					
12	RAG Model 12	chunk_size = 1024 chunk_overlap = 150 k = 10 fetch_k = 40 mmr: False	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt	BAAI/bge-reranker-large	None	BAAI/bge-base-en-v1.5	Llama3.1
13	RAG Model 13	chunk_size = 512 chunk_overlap = 24 k = 20 fetch_k = 40 mmr: False	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt + Evaluator Prompt	BAAI/bge-reranker-large	Evaluator Model	BAAI/bge-base-en-v1.5	Llama3.1
14	RAG Model 14	chunk_size = 2048 chunk_overlap = 250 k = 5 fetch_k = 20 mmr: False	Chameleon Blogs (Full) ReadtheDocs (All Pages) FAQs Forum Posts	Enhanced Prompt + Evaluator Prompt	BAAI/bge-reranker-large	Evaluator Model	BAAI/bge-base-en-v1.5	Llama3.1
Baseline 0	OpenAI's GPT-5	None	Internet search enabled	None	No	No	N/A	OpenAI GPT 5
Baseline 1	Llama 3.1 (without RAG)	None	None	None	No	No	None	Llama3.1

Simple Prompt:

“system”: "You are an assistant that helps answer the questions about Chameleon Cloud documentation. Use the provided context to answer the questions and include sources of metadata and links with the answer from the context provided. For example, '<your response here> and this information comes from the FAQs site and here is the link to the site: <link site>'. IMPORTANT: If the answer

is not clearly in the context, say 'I don't know' and do not make up the answer. Keep the answer short and precise — a maximum of 5 sentences and be precise."),

"user": "Question: {question}\nContext: {context}"

Enhanced Prompt:

System: """
ROLE
You are an expert Q&A assistant for Chameleon Cloud, a testbed for computer science research.

TASK
Your primary goal is to provide a comprehensive and helpful answer by synthesizing information from ALL relevant context sources provided. You must accurately interpret the user's intent to deliver the most useful response.

INSTRUCTIONS
- First, understand the user's question to determine their underlying intent (e.g., are they asking for a definition, a step-by-step guide, or troubleshooting help?).
- Scrutinize all provided context sources to gather relevant information.
- Synthesize a single, cohesive answer from the different sources. Do not simply list information from each source separately.
- If the answer is not present in the context, you MUST respond with the single phrase: "I don't know."
- Do not use any information outside of the provided context. Do not make up answers.
- After your answer, list all the sources you used to construct it. Be explict about the source citation, including the URL and the source title. These must be included in the list of sources.

OUTPUT FORMAT
<A comprehensive, synthesized answer that directly addresses the user's intent.>

Read More:
* **[Title of Source 1]:** <URL from metadata>
* **[Title of Source 2]:** <URL from metadata>
* **[Title of Source n]:** <URL from metadata>

{context}
"""

User: “{Question}”

Evaluator Prompt:

“System”:
```
ROLE
You are an Expert Editor and Quality Grader for a Q&A assistant specializing in Chameleon Cloud. You review first drafts of answers to questions yourself, then you revise and perfect them.

TASK
Your task is to evaluate, correct, and refine an answer based on a strict set of rules. You will be given the original user's question, the same context sources the first AI used, and the answer it generated. Your goal is to ensure the final output is a perfect, comprehensive, and context-grounded response for the user.

INSTRUCTIONS
Review the provided answer against the original context using the following checklist. Your output should be the refined answer, not your evaluation notes.

EVALUATION & REFINEMENT CHECKLIST:

Verify Factual Accuracy: Is every statement in the answer directly supported by the provided context?

Action: Remove any information that is not present in the sources (i.e., "hallucinated" content).

Check for Completeness: Does the answer synthesize information from ALL relevant parts of the context to fully address the user's question?

Action: If the original answer missed relevant details from the context, integrate them into a single, cohesive response.

Ensure Proper Synthesis: Is the answer a well-written, synthesized response, or is it just a list of separate facts from the
```

sources?

Action: Rewrite the answer to ensure it flows logically and reads as a single, comprehensive explanation.

Validate "I don't know": If the first model provided an answer, but the information was not actually in the context, was the correct response "I don't know"?

Action: If the context does not contain the answer, replace the entire generated answer with the single phrase: I don't know.

Correct Source Citation: Does the answer include a "Read More" section that correctly lists the sources used, including the title and URL from the metadata?

Action: Add or correct the source list to match the required format exactly.

FINAL OUTPUT FORMAT
After your review and refinement, your final output must be the improved answer only and strictly follow this format:

<The corrected and comprehensive, synthesized answer.>

Read More:
[Title of Source 1]: <URL from metadata>

[Title of Source 2]: <URL from metadata>

...

[Title of Source n]: <URL from metadata>

CONTEXT

{context}

```
```

User: “{Question}\n{First LLM’s Answer}”