Chameleon Concierge: Retrieval-Augmented Generation (RAG) To Enhance Open Testbed Documentation

Saieda Ali Zada, Marc Richardson (Advisor), Kate Keahey[15] (Advisor)
University of Delaware, University of Chicago, Argonne National Laboratory

References

Good Infrastructure Demands Good Documentation

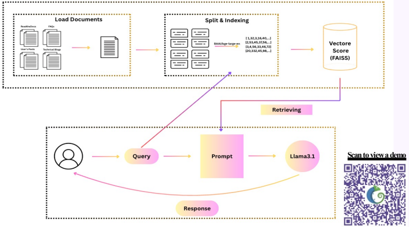
- ❖ Computing infrastructure for open science enables complex, large-scale experiments in computer and domain sciences
- ❖ Experimental design and methodology selection for testbeds requires expertise across multiple technical resource types
- ❖ Researchers need guidance to match their experimental hypotheses with appropriate infrastructure resources, configurations, and methodologies

Where Do Researchers Struggle?

- ❖ **Searching for comprehensive technical solutions** across multiple, disparate documentation sources is a challenge
- ❖ Leads to opening a support ticket or project abandonment**, redirecting infrastructure operators away from other key operations and reducing research impacts**
- ❖ **Solution**: implement a custom **LLM search service** for documentation to generate **accurate and cited responses to natural language queries**

How Can Advances in LLMs and RAG Help?

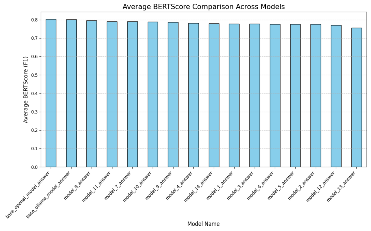
- ❖ Combine conventional (ReadtheDocs) and non-convention (usage data; user tickets) docs for efficient information search to user queries
- ❖ Pull relevant "slices" of information from diverse sources that respond most comprehensively to the user's question
- ❖ Pass along context and sources with question to an LLM to generate a robust response with direct links to sources and up-to-date info



Scan to view a demo

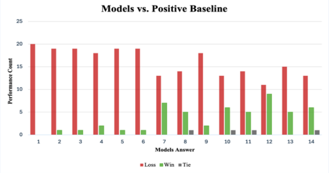
Grading the System's Answers

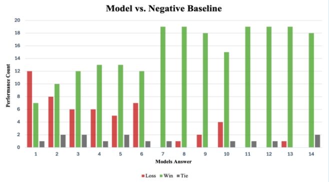
Statistical metrics for textual distances to evaluate quality:



- ❖ Compared RAG model with 20 reference answers to common questions
- ❖ Calculated statistical similarity to compare answers, i.e., BERTScore, but metrics were of limited value for evaluating the system
- ❖ Utilized LLM as a Judge (Claude 3.5 Sonnet) to compare positive baseline (expected best performance), negative baseline (expected to perform worst), and RAG answers (see images on the right)
- ❖ LLM Judge score winning answers by "win", "loss", and "tie" between the baselines and the RAG answers

LLM-as-a-Judge to compare pairwise and select best answer of each match-up:





Insights

Summary:

- ❖ RAG models generated **accurate and cited answers to a variety of user queries**
- ❖ The similarity metrics were not sufficient to determine compare performance; **Judge method provides more meaning evaluation results to determine system quality**
- ❖ Top RAG performance is higher than that of a generic LLM and **comparable to a free-tier proprietary LLM**
- ❖ RAG systems designed around high-quality documentation sources can fill the gap between the researchers' knowledge and limitations of static documentation
- ❖ RAG is not a guaranteed replacement for existing proprietary models, but optimized correctly, one can yield definite benefits

Future work:

- ❖ Enhance data sources by including specialized data (i.e., user ticket data sanitized to remove private data)
- ❖ Explore new generation designs and other evaluation methods through user-provided rankings of answers

The results of this poster were obtained on the Chameleon Testbed funded by the National Science Foundation (Award No. 2431425).
We are grateful to Kexin Pei at the University of Chicago for his helpful comments.

ACM Student Research Competition, SC'25 , St. Louis, MO

saieda@udel.edu