The recent significant advances in the efficacy of large language models (LLMs) have led to a high demand in innovative approaches to using LLMs in educational settings and research (Akpan et al., 2024). Some uses are elaborate such as the creation of a simulated classroom powered by LLMs acting as teacher, assistant and student agents managing to demonstrate interaction patterns like real classrooms between agents and other agents as well as between agents and students (Zhang et al., 2024). Others examined teacher-led interventions that enhance student interactions with LLMs, aiming to optimize the educational benefits of these tools (Kumar et al., 2024).

Large language models have shown potential as personalized tutors able to break down problems and prompt students to come up with the answers as well as adjusting to the needs of individual learners (Chen et al., 2023) (Sonkar et al., 2023). LLMs remain difficult to train with one challenge being ensuring they have updated information to give accurate answers (Chen et al., 2023).

The merging of pre-trained parametric memory (state of the art LLMs) and non-parametric memory, such as a vector database, in retrieval augmented generation (RAG) models provides a method to give LLMs access to updated information without undergoing new training. These models deliver more specific and factual responses (Lewis et al., 2020) and can reduce factual hallucination and increase domain-specific expertise (Chen et al., 2023). This methodology enhances the potential for LLMs to act as AI tutors (Chen et al., 2024), and for RAG pipelines to be utilized to assess data sources. This research examines a methodology for assessing databases in the context of keeping up with rapidly changing cybersecurity threats. Research questions:

Is Reddit an effective source of data for a RAG to address cybersecurity questions? Does chunk size of data affect the accuracy of responses from the LLM?

**Methodology**

**Scrape data from Reddit:**
- Reddit is a potential real-time source of updated cyber threat knowledge. Top 500 posts and their top-voted response were mined from multiple subreddits related to cybersecurity.

**Store in vector database using Chroma:**
- Identifies the relevancy of words to feed the right context to LLM. Database generated for each chunk size.

**Generate 30 questions and ground truths:**
- Developed from threat updates on the Cybersecurity & Infrastructure Security Agency website.

**Feed through Llama RAG pipeline (Llama + Reddit data):**
- Llama answers the 30 questions using the data chunks provided by the database

**Use RAGAs to analyze answer correctness:**
- Llama answers are compared to ground truths.

**Results and Discussion**

Answer correctness average was close to 50% for every chunk size. Some questions yielded better results than others. Reddit is not a great source for a RAG database to provide timely cybersecurity answers considering 50% accuracy is essentially a coin toss. Other potential sources can be evaluated. Chunk size did not have a significant effect on answer correctness in our context though chunk sizes 350 and 4000 did return the highest average answer correctness.

**Answer Correctness by Chunk Size**

<table>
<thead>
<tr>
<th>Chunk Size</th>
<th>Average Answer Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.502507787</td>
</tr>
<tr>
<td>500</td>
<td>0.42961572</td>
</tr>
<tr>
<td>1000</td>
<td>0.39740992</td>
</tr>
<tr>
<td>2000</td>
<td>0.45981484</td>
</tr>
<tr>
<td>4000</td>
<td>0.500473855</td>
</tr>
</tbody>
</table>

The methodology is successful in achieving the goals to answer the research questions. Other sources can be used for the database to determine a source that would be helpful for providing updated threat information. The model can also be used by teachers to vet the effectiveness of teaching materials or the appropriateness of their assessment questions.

**Conclusion**

For an AP Cybersecurity class, some elements such as vulnerabilities and fixes can be discussed in a classroom setting. There could be some activities involving the LLM and how RAG works to optimize it. RAG combines LLMs with retrieval mechanisms to access external knowledge. LLMs: Use deep learning to generate human-like text based on input. Discussion: how can RAG and LLMs be applied to cybersecurity?

Lecture: Discuss the importance of real-time information retrieval in cybersecurity, incident response, threat intelligence.

Examples: Present real-world examples of RAG in cybersecurity tools. Hands-on Exercise: Setting Up the Environment

Activity: Guide students to set up a Jupyter Notebook or Google Colab.

API Access: Provide instructions for accessing a language model API.

The methodology also has potential as a tool for teachers to analyze the suitability of classroom resources or of assessments.

**Acknowledgements**

Thank you to our research mentors Dr. Artiran Pipiali and Kristyls Ruiz-Rohena for the support and guidance in this project. Thanks also to the entire CREDS REU team at UTEP for the opportunity and support.

---

**References**


