# Experiment Report: Evaluating LLM Performance with Zero-Shot vs. Chain-of-Thought Prompting in RAG

Ziad Amr Shamseldein (52-0295)

Hazem Sherif (52-5272)

Youssef Bayoumi (52-11219)

## Methodology

The experiment aimed to compare the effectiveness of two prompting techniques—Zero-Shot and Chain-of-Thought (CoT)—when integrated with a Retrieval-Augmented Generation (RAG) pipeline. The model used was Google's FLAN-T5-base, and the evaluation was conducted on the SQuAD v2 dataset. The RAG pipeline was built using FAISS for efficient retrieval of relevant context from the dataset, and the HuggingFace Transformers library was employed for model inference.

For Zero-Shot prompting, the model was directly queried with questions, while the CoT approach involved breaking down the reasoning process into structured steps (e.g., identifying the question's intent, citing relevant context, and formulating a hypothesis). The evaluation metric was ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which measures the overlap between generated answers and ground-truth references.

## Results

The ROUGE scores revealed that Zero-Shot prompting outperformed Chain-of-Thought prompting in this setup:

Zero-Shot Prompting: Achieved ROUGE-1 of **0.396**, ROUGE-2 of **0.216**, and ROUGE-L of **0.395**.

Chain-of-Thought Prompting: Achieved ROUGE-1 of 0.331, ROUGE-2 of **0.199**, and ROUGE-L of **0.329**.

The higher scores for Zero-Shot prompting suggest that the model performed better when answering questions directly, without the additional reasoning steps imposed by CoT. This could be attributed to the model's inherent ability to generate concise answers when the retrieved context is already highly relevant.

**Limitations**

One key limitation of the experiment is the reliance on a single dataset (SQuAD v2), which may not fully represent the diversity of real-world queries. Additionally, the CoT prompting template used was generic and might not have been optimized for the specific task, potentially leading to suboptimal performance. The evaluation also did not account for the quality of reasoning in CoT responses, as ROUGE primarily measures lexical overlap rather than logical coherence.

Another limitation is the computational overhead of CoT prompting, which requires more tokens and longer processing times due to the step-by-step reasoning. This was evident in the slower response times during interactive testing. Furthermore, the experiment did not explore fine-tuning the model specifically for CoT tasks, which could have improved its performance.

## Suggested Improvements

To enhance the experiment, future work could involve designing more tailored CoT prompts that align closely with the dataset's question-answer pairs. Fine-tuning the model on CoT-style reasoning tasks could also improve its ability to generate structured and accurate reasoning steps. Additionally, incorporating other evaluation metrics, such as BLEU or human judgment, could provide a more comprehensive assessment of answer quality beyond lexical overlap.

Exploring hybrid approaches, where Zero-Shot and CoT are dynamically selected based on question complexity, might yield better results. For instance, simpler questions could use Zero-Shot, while complex or multi-step questions could leverage CoT. Finally, testing on a broader range of datasets, including those requiring multi-hop reasoning, would validate the generalizability of the findings.

## Conclusion

The experiment demonstrated that Zero-Shot prompting can be more effective than Chain-of-Thought prompting in a RAG pipeline for question-answering tasks, as measured by ROUGE scores. However, the results highlight the need for further optimization of CoT techniques and evaluation metrics to better capture the nuances of reasoning-based responses. Future iterations could focus on refining prompts, incorporating model fine-tuning, and expanding the scope of evaluation to ensure robust and scalable performance across diverse use cases.