Frankfurt University
of Applied Sciences

– Faculty of Computer Science and Engineering –

Evaluating the Role of Chunk Size in Retrieval-Augmented Generation (RAG)

Thesis submitted in order to obtain the academic degree

Bachelor of Science (B.Sc.)

submitted on 21. January 2024 by

Md Jewel Rana

Student ID: 1398135

First Supervisor	:	Prof. Dr. Egbert Falkenberg
Second Supervisor	:	Prof. Dr. Eicke Godehardt

Declaration

Hereby I assure that I have written the presented thesis independently and without any third party help and that I have not used any other tools or resources than the ones mentioned/referred to in the thesis.

Any parts of the thesis that have been taken from other published or yet unpublished works in terms of their wording or meaning are thoroughly marked with an indication of the source.

All figures in this thesis have been drawn by myself or are clearly attributed with a reference.

This work has not been published or presented to any other examination authority before. I am aware of the importance of the affidavit and the consequences under examination law as well as the criminal consequences of an incorrect or incomplete affidavit.

Frankfurt, 21. January 2024

Md Jewel Rana

Contents

List of Figures

Abstract

Chapter 1

Introduction

1.1 Background and Context

Introduction to Large Language Models

Large Language Models (LLMs) are advanced AI systems capable of analyzing and generating human language. They can intelligently analyze text to produce coherent responses and execute various language-related tasks. In the business world, LLMs can unlock valuable insights from vast amounts of text data, automate time-consuming tasks, and enhance customer experiences through more personalized and efficient interactions.[4]

Limitations of Standalone Large Language Models

However, despite their power, LLMs face critical limitations when they are used as a standalone system. In many cases, LLMs generate responses that sound plausible but are factually incorrect. This phenomenon is known as hallucination. This occurs because LLMs rely on pre-trained data, which may not contain all the necessary knowledge to accurately answer domain-specific questions. [1][9] Additionally, pre-trained LLMs are limited by the static knowledge they acquired during training. They cannot access recent developments or specialized domain knowledge that wasn't included in their training data. [1][9] Furthermore, many real-world applications, such as legal, medical, or technical fields, require domain-specific expertise. LLMs alone may lack the depth that may necessary to handle such specialized queries effectively. [1] Moreover, in question-answering tasks, users often expect accurate, up-to-date, and contextually relevant responses. [9]

Introduction to the Retrieval-Augmented Generation

To overcome these limitations, a new paradigm called Retrieval-Augmented Generation (RAG) has been introduced. RAG is a framework which is designed to enhance the performance of LLMs by incorporating an external mechanism. Instead of completely depending on the pre-trained knowledge base within the model, RAG retrieves relevant document chunks from external knowledge bases based on semantic similarity calculations. This enables LLMs to access and incorporate up-to-date, domain-specific information into their responses. By referencing external knowledge, RAG addresses a significant limitation of traditional LLMs: their tendency to generate factually incorrect content when they are faced with queries that are not in their training data. The retrieved information from the external knowledge base grounds the LLMs's responses, reducing the problem of hallucinations. Due to this integration, RAG has become a widely adopted technology, particularly in advancing chatbot capabilities and improving the practical applicability of LLMs in real-world applications.[1]

Challenges in RAG Implementation

However, with the development of RAG technique, new challenges have emerged. A reliable retrieval pipeline is crucial for building effective RAG-based applications. The quality of the final answer highly depends on the relevance of the retrieved text to the user's query. If the retriever fails to identify relevant information from the knowledge bases, the LLMs may generate inaccurate or misleading responses.[5] One of the most crucial factors that directly influences the performance of semantic retrieval is chunk size.[3] In the chunking process, a large corpus of text data is divided into smaller and semantically meaningful units.[10] Smaller chunks contain more focused and context-specific information, improving the relevance of retrieved results. However, they may lack the broader context necessary to answer certain queries. Conversely, larger chunks include more context, some of which may not be relevant to the specific query. This can lead to less accurate similarity scores and potentially less useful results for RAG.[3]

1.2 Motivation

The evaluation of the impact of chunk size on RAG systems is highly significant because it directly affects the performance and reliability of applications such as chatbots and QA systems. In RAG-based systems, LLMs are enhanced by integrating external knowledge sources which enable them to generate more accurate and contextually relevant responses. [6]
However, the effectiveness of this integration can be influenced by the chunk size, or how the corpus data is segmented. [1] In real-world applications like chatbots, the precision and relevance of responses are crucial. Improper chunking may lead to the retrieval of irrelevant or incomplete information which results in responses that are factually incorrect or lack coherence. [9] For example, in customer support scenarios, a chatbot must access specific sections of the knowledge base to accurately answer user queries. Because of large chunk size the system may retrieve unnecessary information which may overwhelm the user with irrelevant details. [3] On the other hand, if the chunks are too small, critical context may be missed by the retrieval system which may lead to incomplete answers. [6] Effective chunking ensures that the system can extract relevant documents without being misled by irrelevant content. [9] This precision affects the system's ability to provide more accurate and concise answers. [1]

Therefore, understanding and optimizing chunk size is essential for improving the performance of RAG-based applications. By ensuring optimal chunk sizes, developers can improve the accuracy of information retrieval and build efficient AI-driven communication tools. [6]

1.3 Research Problem

1.3.1 Main Research Question

The main question of this research is to evaluate the influence of chunk size on the accuracy, relevance, and efficiency of responses generated by RAG models in open-domain question answering.

1.3.2 Supporting Questions

In order to thoroughly address this primary question, the following supporting questions will also be explored:

- **Impact on Factual Accuracy:** How does chunk size affect the factual accuracy of RAG-generated answers?
- **Mitigating Hallucinations:** Are shorter or longer chunk size more effective in minimizing hallucinations in generated responses?
- **Optimal Chunk size:** What is the optimal chunk to balance retrieval efficiency and response quality in RAG models?

- **Performance Metrics:** Does chunk size impact the generation speed and real-time applicability of RAG-based systems?
- **Context vs. Irrelevance:** How do larger chunks compare to smaller ones in contributing irrelevant details that may affect response quality?

1.4 Objectives

The objectives of this thesis are designed to systematically investigate the role of chunk size on the performance of RAG based systems. Each of these objectives is specifically tailored to assess a crucial aspect of model performance:

1.4.1 Evaluate Chunk size and Accuracy

One of the primary objectives of this thesis is to assess how different chunk sizes, such as short, medium, and long, affect the quality and factual accuracy of RAG-generated responses. Chunk size directly impacts the contextual information available to the model during generation.[3] This objective focuses on quantifying the effects of chunk size on accuracy through rigorous experimentation. In the evaluation, metrics such as F1-score and ROUGE will be employed to assess the similarity of generated answers to verified reference answers.

1.4.2 Analyze Hallucination Rates

RAG systems, like many other generative models often produce hallucinations—responses containing fabricated or unsupported or irrelevant information. These hallucinations can undermine the trustworthiness and usability of the system.[7] This objective intends to evaluate and classify the prevalence of hallucination as a function of document length. Hallucinations will be categorized into two classes:

- **Intrinsic hallucinations**, these errors may arise from the generative model, often resulting from pre-trained biases or insufficient contextual understanding.[7]
- **Extrinsic hallucinations**, these errors are a consequence of the system's reliance on incomplete, irrelevant, or noisy information extracted from its knowledge base.[7] The system's inability to accurately identify and filter out such information can lead to the generation of misleading outputs.

1.4.3 Assess Retrieval and Generation Speed

Efficiency is a vital factor in the development of RAG based systems, most importantly in real-time applications. Different chunk sizes impact the speed of retrieval and response generation.[8] The aim of this objective is, to measure the retrieval times and the generation speeds for each chunk size to find out the trade-offs between speed and quality, that are highly required in real-time applications. These assessments will provide insights into the feasibility of RAG systems for real-time use cases.

1.4.4 Provide Practical Recommendations

Leveraging the experimental outcomes, this thesis will provide recommendations for the optimization of document length selection in RAG based QA System. These guidelines will be applicable to a variety of use cases, including customer support, and QA systems.

1.5 Thesis Structure

Provide an overview of how the thesis is organized.

Chapter 2

Literature Review

2.1 Foundational Studies

In this section the key foundational works that form the basis of RAG systems will be explained briefly. Focus will be on two particularly relevant studies. These studies provide a strong foundation for understanding the core concepts and challenges in RAG research.

Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation

Zijie Zhong et al.[11] introduces the Mix-of-Granularity (MoG) approach. In this approach chunk sizes are dynamically adjusted based on the query requirements to optimize RAG performance. This study highlights the trade-offs between retrieval efficiency and response quality. A hierarchical retrieval framework is used to achieve MoG that combines smaller chunks for precision with larger chunks for context, which dynamically balances granularity to suit different tasks. This study establishes a strong evidence that chunk size directly impacts retrieval speed, hallucination rates, and the overall factual accuracy of generated responses. This research provides critical insights into the effects of chunk granularity, that forms a theoretical basis for examining chunk size systematically.

Financial Report Chunking for Effective Retrieval-Augmented Generation

Another Study [2], which was conducted by Antonio Jimeno Yepes et al. that experimented chunking strategies specifically in the context of financial documents. Their study highlights the impact of chunk size on RAG systems, particularly while dealing with complex, structured data. They propose methodologies for chunking financial reports into semantically coherent segments, which demonstrate that appropriate chunk sizes improve retrieval relevance and generative accuracy while reducing hallucinations. This work also highlights the challenges in processing lengthy financial documents, that emphases the need for tailored chunking strategies to ensure performance in domain-specific applications.

These foundational studies lay the groundwork for understanding how chunk size affects RAG models, particularly in diverse and complex domains such as open-domain QA and structured document retrieval.

2.2 Related Work

This section explores relevant literature that intersects with the research focus, which offer insights into document retrieval, generative systems, and the specific influence of document characteristics like length or chunk size.

2.3 Gap Analysis

Chapter 3

Theoretical Background

3.1 RAG architecture

3.2 Sentence Embeddings

3.3 Sentence Transformer

3.4 Vector Search

3.4.1 Euclidean Distance

3.4.2 Cosine Similarity

3.5 Chunking in RAG

3.5.1 Introduction to Chunking

3.5.2 Chunking Strategies

3.5.2.1 Fixed-Size Chunking

3.5.2.2 Recursive-Based Chunking

3.5.2.3 Document-Based Chunking

3.5.2.4 Semantic Chunking

3.5.2.5 Hybrid Chunking

3.5.3 Impact of chunk size on RAG

Impact on Information Retrieval

Impact on Generation

Impact on retrieval time and response time

3.6 Metrics and Evaluation

Chapter 4

Methodology

Chapter 5

Experiments and Results

5.1 Experiment 1 : Chunk size and Factual Accuracy

5.2 Experiment 2 : Hallucination Analysis

5.3 Experiment 3 : Retrieval and Generation Speed

Chapter 6

Discussion

6.1 Analysis of Results

6.2 Insights on Chunk Size

6.3 Practical Implications

6.4 Limitations

Chapter 7

Conclusion

7.1 Summary of Findings

7.2 Contributions

7.3 Future Work

List of Abbreviations

AI Artificial Intelligence. 2, 3

LLMs Large Language Models. 2, 3

MoG Mix-of-Granularity. 5

QA Questions & Answers. 3–5

RAG Retrieval-Augmented Generation. 2–5

ROUGE Recall-Oriented Understudy for Gisting Evaluation.. 4

Bibliography

[1] Yunfan Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey." In: *arXiv preprint arXiv:2301.12730* (2023). URL: https://arxiv.org/abs/2301.12730.

[2] Antonio Jimeno Yepes et al. "Financial Report Chunking for Effective Retrieval-Augmented Generation." In: *ArXiv preprint* (2024). URL: https://unstructured.io.

[3] Lam Hoang. *How Chunk Sizes Affect Semantic Retrieval Results*. Mar. 11, 2024. URL: https://ai.plainenglish.io/investigating-chunk-size-on-semantic-results-b465867d8ca1.

[4] Patricia Kelbert, Dr. Julien Siebert, Lisa Jöckel. *Was ist ein Large Language Model (LLM)?* Dec. 12, 2023. URL: https://www.iese.fraunhofer.de/blog/large-language-models-ki-sprachmodelle/ (visited on 11/22/2024).

[5] Ishneet Sukhvinder Singh et al. "ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems." In: *arXiv preprint arXiv:2410.19572* (2024). DOI: 10.48550/arXiv.2410.19572. URL: https://doi.org/10.48550/arXiv.2410.19572.

[6] Ishneet Sukhvinder Singh et al. "ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems." In: *arXiv preprint arXiv:2410.19572* (2024). DOI: 10.48550/arXiv.2410.19572. URL: https://doi.org/10.48550/arXiv.2410.19572.

[7] DL Staff. "Understanding Intrinsic and Extrinsic Hallucinations in NLP." In: *Communications of the ACM* 66.12 (2023), pp. 202–215. URL: https://dl.acm.org/doi/pdf/10.1145/3632410.3633297.

[8] LlamaIndex Team. *Evaluating the Ideal Chunk Size for a RAG System Using LlamaIndex*. Accessed: 2024-11-23. 2023. URL: https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5.

[9] Hao Yu et al. "Evaluation of Retrieval-Augmented Generation: A Survey." In: *arXiv preprint arXiv:2301.07297* (2023). URL: https://arxiv.org/abs/2301.07297.

[10] Tong Zhang, Fred Damerau, and David Johnson. "Text chunking based on a generalization of Winnow." In: *Journal of Machine Learning Research* 2.4 (2002).

[11] Zijie Zhong et al. "Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation." In: *ArXiv preprint* (2024). URL: https://arxiv.org/abs/2406.00456.