Eser Adı: Editorial: Large language models in work and business

Atıflar

1. Pawlik, Lukasz. "LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems" APPLIED SCIENCES-BASEL (2025). https://doi.org/10.3390/app152010886

Atıf 1

Pawlik, Lukasz. "LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems" APPLIED SCIENCES-BASEL (2025). https://doi.org/10.3390/app152010886

A.1. Yayının Ünvan Sayfası (kitap, dergi, vb.)

applied sciences

Impact Factor 2.5
CiteScore 5.5

ISSN 2076-3417

Bioactive Compounds from Porphyra umbilicalis: Implications for Human Nutrition

Volume 15 · Issue 20 · October-2 2025

MDPI

Journals Topics Information Author Services Initiatives About

Sign In / Sign Up Submit

You can access the new MDPI.com website here. Explore and share your feedback with us.

Search for Articles: Title / Keyword Author / Affiliation / Email Applied Sciences All Article Types Search Advanced

Journals / Applied Sciences / Volume 15 / Issue 20 / 10.3390/app152010886

applied sciences

Submit to this Journal

Review for this Journal

Propose a Special Issue

Article Menu

Academic Editor

Michal Ptaszynski

Recommended Articles

Related Info Link

More by Author Links

Article Views 3007

Citations 1

Table of Contents

Abstract

Introduction

Order Article Reprints

Share

Help

Cite

Discuss in SciProfiles

Open Access Article

LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems

by Lukasz Pawlik

Department of Information Systems, Kielce University of Technology, 7 Tysiąclecia Państwa Polskiego Ave., 25-314 Kielce, Poland

Appl. Sci. **2025**, *15*(20), 10886; https://doi.org/10.3390/app152010886

Submission received: 22 September 2025 / Revised: 30 September 2025 / Accepted: 9 October 2025 / Published: 10 October 2025

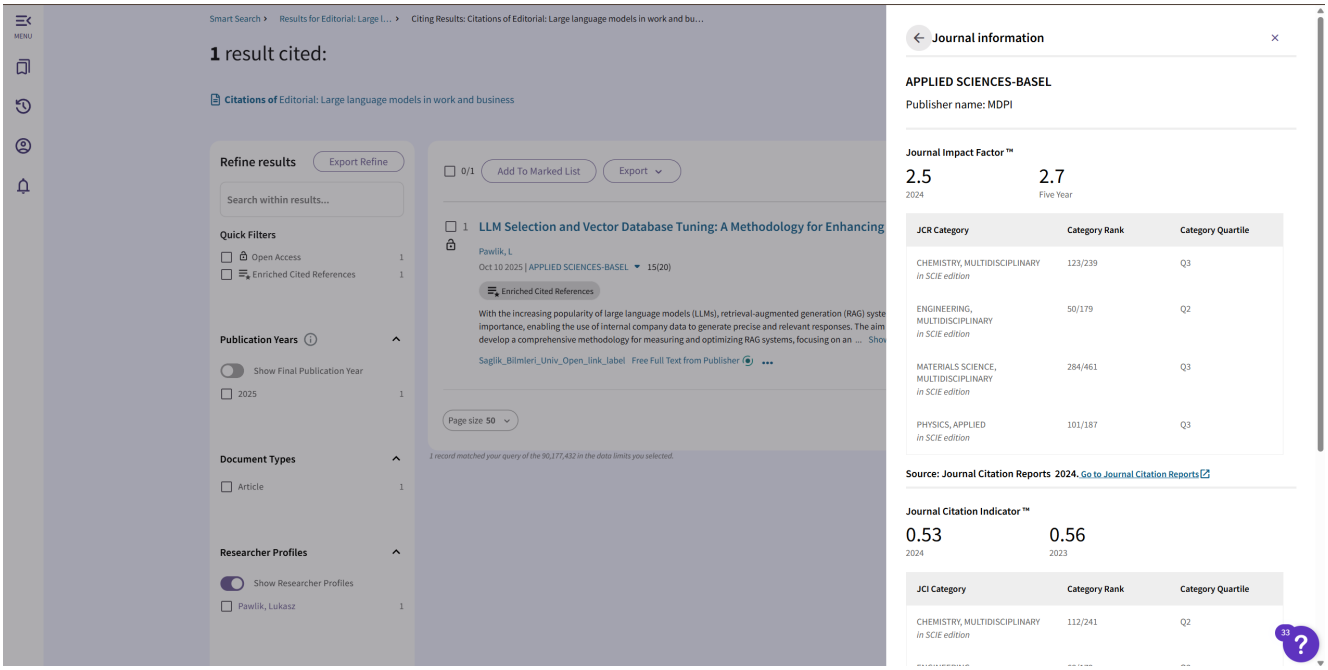
Download Browse Figures Versions Notes

Abstract

With the increasing popularity of large language models (LLMs), retrieval-augmented generation (RAG) systems are gaining importance, enabling the use of internal company data to generate precise and relevant responses. The aim of this study was to develop a comprehensive methodology for measuring and optimizing RAG systems, focusing on analyzing the impact of key parameters such as chunk size, vector embedding models, and LLM selection on system effectiveness. Experiments were conducted on a RAG system using a large biographical dataset, stored in a Qdrant vector database, allowing for in-depth analysis in the context of long text data. The results indicated that optimizing RAG systems necessitates considering various factors, including LLM context window size, computational power, and processing costs. The selection of optimal parameters and LLM is a trade-off between response quality, computational cost, and hardware limitations. This study provides practical guidance for engineers and researchers working on improving RAG-based systems, enabling informed decisions regarding RAG system configuration in various business contexts.

Keywords: retrieval-augmented generation; vector database; large language model

1. Introduction


Smart Search › Results for Editorial: Large l… › Citing Results: Citations of Editorial: Large language models in work and bu…
1 result cited:
Citations of Editorial: Large language models in work and business
Refine results
Export Refine
Search within results...
Quick Filters
Open Access 1
Enriched Cited References 1
Publication Years
Show Final Publication Year
2025 1
Document Types
Article 1
Researcher Profiles
Show Researcher Profiles
Pawlik, Lukasz 1
0/1
Add To Marked List
Export
1 LLM Selection and Vector Database Tuning: A Methodology for Enhancing
Pawlik, L
Oct 10 2025 | APPLIED SCIENCES-BASEL 15(20)
Enriched Cited References
With the increasing popularity of large language models (LLMs), retrieval-augmented generation (RAG) syste
importance, enabling the use of internal company data to generate precise and relevant responses. The aim
develop a comprehensive methodology for measuring and optimizing RAG systems, focusing on an ... Show
Saglik_Bilmleri_Univ_Open_link_label Free Full Text from Publisher
Page size 50
1 record matched your query of the 90,177,432 in the data limits you selected.
Journal information
APPLIED SCIENCES-BASEL
Publisher name: MDPI
Journal Impact Factor ™
2.5
2024
2.7
Five Year
JCR Category
Category Rank
Category Quartile
CHEMISTRY, MULTIDISCIPLINARY
in SCIE edition
123/239
Q3
ENGINEERING, MULTIDISCIPLINARY
in SCIE edition
50/179
Q2
MATERIALS SCIENCE, MULTIDISCIPLINARY
in SCIE edition
284/461
Q3
PHYSICS, APPLIED
in SCIE edition
101/187
Q3
Source: Journal Citation Reports 2024. Go to Journal Citation Reports
Journal Citation Indicator ™
0.53
2024
0.56
2023
JCI Category
Category Rank
Category Quartile
CHEMISTRY, MULTIDISCIPLINARY
in SCIE edition
112/241
Q2
33

A1. Eserin Başlık Sayfası

applied sciences

MDPI

Article

LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems

Lukasz Pawlik

Department of Information Systems, Kielce University of Technology, 7 Tysiąclecia Państwa Polskiego Ave., 25-314 Kielce, Poland; lpawlik@tu.kielce.pl

Abstract

With the increasing popularity of large language models (LLMs), retrieval-augmented generation (RAG) systems are gaining importance, enabling the use of internal company data to generate precise and relevant responses. The aim of this study was to develop a comprehensive methodology for measuring and optimizing RAG systems, focusing on analyzing the impact of key parameters such as chunk size, vector embedding models, and LLM selection on system effectiveness. Experiments were conducted on a RAG system using a large biographical dataset, stored in a Qdrant vector database, allowing for in-depth analysis in the context of long text data. The results indicated that optimizing RAG systems necessitates considering various factors, including LLM context window size, computational power, and processing costs. The selection of optimal parameters and LLM is a trade-off between response quality, computational cost, and hardware limitations. This study provides practical guidance for engineers and researchers working on improving RAG-based systems, enabling informed decisions regarding RAG system configuration in various business contexts.

Keywords: retrieval-augmented generation; vector database; large language model

check for updates

Academic Editor: Michal Ptaszynski

Received: 22 September 2025
Revised: 30 September 2025
Accepted: 9 October 2025
Published: 10 October 2025

Citation: Pawlik, L. LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems. *Appl. Sci.* **2025**, *15*, 10886. https://doi.org/10.3390/app152010886

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, large language models (LLMs) have revolutionized business approaches, introducing new possibilities in process automation, data analysis, and customer interaction. Thanks to advanced natural language processing algorithms, LLMs can generate answers to questions, create personalized recommendations, and support users in decision-making. In various fields, such as medicine, law, and education, LLMs offer tools that can significantly streamline these processes by providing precise and relevant information.

However, their effectiveness is often limited by the lack of access to internal company data, which prevents the full utilization of these technologies in the context of specific organizational needs and data. This limitation hinders the ability of organizations to leverage LLMs for tasks requiring specific internal knowledge, leading to inaccurate or irrelevant responses, reduced efficiency, and missed opportunities for data-driven decision-making [1–3].

The solution to this problem is the application of retrieval-augmented generation (RAG) technology, which enables the integration of external data sources with the company's internal information resources. RAG combines the text generation capabilities of language models with access to external and internal data sources, allowing for the use of specific organizational information resources to generate more precise and relevant

A1. Eserde atıf yapılan sayfa(lar)

applied sciences

MDPI

Article

LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems

Lukasz Pawlik

Department of Information Systems, Kielce University of Technology, 7 Tysiąclecia Państwa Polskiego Ave., 25-314 Kielce, Poland; lpawlik@tu.kielce.pl

Abstract

With the increasing popularity of large language models (LLMs), retrieval-augmented generation (RAG) systems are gaining importance, enabling the use of internal company data to generate precise and relevant responses. The aim of this study was to develop a comprehensive methodology for measuring and optimizing RAG systems, focusing on analyzing the impact of key parameters such as chunk size, vector embedding models, and LLM selection on system effectiveness. Experiments were conducted on a RAG system using a large biographical dataset, stored in a Qdrant vector database, allowing for in-depth analysis in the context of long text data. The results indicated that optimizing RAG systems necessitates considering various factors, including LLM context window size, computational power, and processing costs. The selection of optimal parameters and LLM is a trade-off between response quality, computational cost, and hardware limitations. This study provides practical guidance for engineers and researchers working on improving RAG-based systems, enabling informed decisions regarding RAG system configuration in various business contexts.

Keywords: retrieval-augmented generation; vector database; large language model

check for updates

Academic Editor: Michal Ptaszynski

Received: 22 September 2025
Revised: 30 September 2025
Accepted: 9 October 2025
Published: 10 October 2025

Citation: Pawlik, L. LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems. *Appl. Sci.* **2025**, *15*, 10886. https://doi.org/10.3390/app152010886

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, large language models (LLMs) have revolutionized business approaches, introducing new possibilities in process automation, data analysis, and customer interaction. Thanks to advanced natural language processing algorithms, LLMs can generate answers to questions, create personalized recommendations, and support users in decision-making. In various fields, such as medicine, law, and education, LLMs offer tools that can significantly streamline these processes by providing precise and relevant information.

However, their effectiveness is often limited by the lack of access to internal company data, which prevents the full utilization of these technologies in the context of specific organizational needs and data. This limitation hinders the ability of organizations to leverage LLMs for tasks requiring specific internal knowledge, leading to inaccurate or irrelevant responses, reduced efficiency, and missed opportunities for data-driven decision-making [1–3].

The solution to this problem is the application of retrieval-augmented generation (RAG) technology, which enables the integration of external data sources with the company's internal information resources. RAG combines the text generation capabilities of language models with access to external and internal data sources, allowing for the use of specific organizational information resources to generate more precise and relevant

information [4,5]. RAG optimization can significantly improve the accuracy and efficiency of systems, providing contextually relevant responses based on unique company data [6].

Despite the proliferation of RAG optimization guides and frameworks, a critical gap remains, as these resources often address the selection and tuning of components (e.g., chunk size, embedding models, LLMs) in isolation. There is a lack of a unified, comprehensive methodology that systematically evaluates and optimizes the combined impact of vector database configuration (via chunking and embedding selection) and LLM selection on an RAG system's effectiveness, particularly in practical settings involving long-form internal documents. Addressing this research gap, this article presents a novel and comprehensive methodology for measuring and optimizing RAG systems. The work focuses on analyzing the synergistic impact of key parameters, including chunk size, vector embedding models, and LLM selection, on overall system effectiveness. The study is based on an experimental RAG system using a large biographical dataset, allowing for in-depth analysis and optimization in the context of long text data. The article details the system architecture, the response generation process, and the evaluation methodology, including the use of LLMs to evaluate results, thereby providing practical guidance on selecting optimal, integrated parameters (chunk size, vector embedding models, LLMs) for improved accuracy and efficiency, particularly within the context of long text data and internal company knowledge.

The structure of this article is as follows: Section 2 presents a review of related work in the field of retrieval-augmented generation (RAG) systems. Section 3 describes the dataset used in the study. Section 4 presents the architecture of the research environment. Section 5 discusses the methodology for measuring the effectiveness of the RAG system. Section 6 describes the methodology for optimizing the parameters of the RAG system. Section 7 analyzes the results of the experiments conducted. Section 8 summarizes the key findings of the research. Section 9 discusses the limitations of the work, and Section 10 proposes directions for future research.

2. Related Work

In recent years, we have observed a dynamic development of retrieval-augmented generation (RAG) technology, which is reflected in the growing number of scientific publications. A key element of these systems is vector databases, which enable efficient storage and retrieval of semantic information. In the era of large datasets, the optimization of vector database algorithms becomes crucial [7]. Researchers are analyzing various approaches to approximate nearest neighbor search (ANNS), such as hashing, tree-based, graph-based, and quantization methods. Hashing methods offer speed at the cost of accuracy, while tree-based methods provide better accuracy but are more computationally demanding [8]. In the context of specific applications, such as the analysis of atmospheric motion vector data, precise processing and evaluation of vector data in large-scale data processing systems are crucial [9].

Vector database management techniques constitute another significant area of research, focusing on the challenges associated with feature vector processing, such as the ambiguous concept of semantic similarity, large vector sizes, and the difficulty in handling hybrid queries. These studies provide an overview of query processing, storage and indexing, query optimization, and existing vector database systems, including vector quantization and data compression techniques [10]. The issue of software testing in the context of vector databases cannot be overlooked, where generating test data, defining patterns of correct results, and evaluating tests are crucial for ensuring greater reliability and trust in vector database systems [11]. In practical applications, such as question-answering systems in science, vector databases are used to create reliable information systems that provide

A1. Kaynakça Sayfası

5 returns 5 chunks). For each returned chunk, the person's "full name" is read. A metadata query is performed to find the remaining chunks for that person. Chunks are combined into a single entity for each person. Combining vector search with metadata filtering (hybrid search). Returning responses in the form of biographical essays with metadata, in a number consistent with the k-limit parameter.

- Retrieved chunks evaluation. Biographical essays are evaluated by the Evaluate LLM, based on the expected vector response, using accuracy as the metric. The vector DB search results are stored in the Relational DB for further use in the Answer LLM response generation process.
- Answer generation. Biographical essays for returned persons are passed to the Answer LLM, which generates a coherent and contextually relevant response to the user query.
- LLM response evaluation. Responses returned by the answer LLM are evaluated by the evaluation LLM, based on the expected final response (LLM), using accuracy as the metric. This evaluation ensures the accuracy of the final responses before being returned to the user.

Appendix A.3. RAG System Configuration Parameters

The configuration options for all programs are presented in Table A2. For the RAG ingestor program, elements relevant to data loading are configured, such as collection name, chunk size, and vector embedding model. For the RAG assistant and RAG Evaluator, the collection name, which determines the vector embedding model, is specified, and the LLM model and k-limit are reconfigurable.

Table A2. Program configuration.

Parameter	RAG Ingestor	RAG Assistant	RAG Evaluator
Collection name	Yes	Yes	Yes
Chunk size	Yes	Not applicable	Not applicable
Vector model	Yes	Same as collection	Same as collection
LLM model	Not applicable	Yes	Yes
k-limit	Not applicable	Yes	Yes

References

1. Gupta, S.; Ranjan, R.; Singh, S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv* **2024**, arXiv:2410.12837.
2. Şakar, T.; Emekci, H. Maximizing RAG efficiency: A comparative analysis of RAG methods. *Nat. Lang. Process.* **2025**, *31*, 1–25. [CrossRef]
3. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997. [CrossRef]
4. Seker, S.E. Editorial: Large language models in work and business. *Front. Artif. Intell.* **2024**, *7*, 1516832. [CrossRef]
5. Bharathi Mohan, G.; Prasanna Kumar, R.; Vishal Krishh, P.; Keerthinathan, A.; Lavanya, G.; Meghana, M.K.U.; Sulthana, S.; Doss, S. An analysis of large language models: Their impact and potential applications. *Knowl. Inf. Syst.* **2024**, *66*, 5047–5070. [CrossRef]
6. Ferraris, A.F.; Audrito, D.; Caro, L.D.; Poncibò, C. The architecture of language: Understanding the mechanics behind LLMs. *Camb. Forum AI Law Gov.* **2025**, *1*, e11. [CrossRef]
7. Zhu, Z. Strategies for Improving Vector Database Performance through Algorithm Optimization. *Sci. J. Technol.* **2025**, *7*, 138–144. [CrossRef]
8. Han, Y.; Liu, C.; Wang, P. A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge. *arXiv* **2023**, arXiv:2310.11703. [CrossRef]
9. Lee, E.; Todling, R.; Karpowicz, B.M.; Jin, J.; Sewnath, A.; Park, S.K. Assessment of Geo-Kompsat-2A Atmospheric Motion Vector Data and Its Assimilation Impact in the GEOS Atmospheric Data Assimilation System. *Remote Sens.* **2022**, *14*, 5287. [CrossRef]
10. Pan, J.J.; Wang, J.; Li, G. Vector Database Management Techniques and Systems. In Proceedings of the Companion of the 2024 International Conference on Management of Data, SIGMOD/PODS '24, Santiago, Chile, 9–14 June 2024; pp. 597–604. [CrossRef]