Search Results for ABSTRACT

From Files: 2312.10997v5.pdf, 2402.16893v1.pdf, 2405.07437v2.pdf, 2410.15944v1.pdf, 3626772.3657957.pdf, 3644815.3644945.pdf, A_Novel_Linearly_Complex_Extended_Signed_Response- (1).pdf, A_Novel_Linearly_Complex_Extended_Signed_Response-.pdf, arxiv.pdf

---

2405.07437v2.pdf

Evaluation of Retrieval-Augmented Generation: A Survey

Abstract

Retrieval-Augmented Generation (RAG) has recently gained traction in natural language processing. Numerous studies and real-world applications are leveraging its ability to enhance generative models through external informa- tion retrieval. Evaluating these RAG systems, however, poses unique challenges due to their hybrid structure and reliance on dynamic knowledge sources. To better understand these challenges, we conduct A Unified Evaluation Process of RAG (Auepora) and aim to provide a comprehensive overview of the evaluation and benchmarks of RAG systems. Specifically, we examine and compare several quantifiable metrics of the Retrieval and Generation components, such as rele- vance, accuracy, and faithfulness, within the current RAG benchmarks, encom- passing the possible output and ground truth pairs. We then analyze the various datasets and metrics, discuss the limitations of current benchmarks, and suggest potential directions to advance the field of RAG benchmarks.

2402.16893v1.pdf

The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)

Abstract

Retrieval-augmented generation (RAG) is a powerful technique to facilitate language model with proprietary and private data, where data privacy is a pivotal concern. Whereas extensive research has demonstrated the privacy risks of large language models (LLMs), the RAG tech- nique could potentially reshape the inherent behaviors of LLM generation, posing new pri- vacy issues that are currently under-explored. In this work, we conduct extensive empiri- cal studies with novel attack methods, which demonstrate the vulnerability of RAG systems on leaking the private retrieval database. De- spite the new risk brought by RAG on the re- trieval data, we further reveal that RAG can mitigate the leakage of the LLMs training data. Overall, we provide new insights in this paper for privacy protection of retrieval- augmented LLMs, which benefit both LLMs and RAG systems builders. Our code is avail- able at https://github.com/phycholosogy/RAG- privacy .

2410.15944v1.pdf

Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report

Abstract

This paper presents an experience report on the develop- ment of Retrieval Augmented Generation (RAG) systems using PDF documents as the primary data source. The RAG architecture combines generative capabilities of Large Language Models (LLMs) with the preci- sion of information retrieval. This approach has the potential to redefine how we interact with and augment both structured and unstructured knowledge in generative models to enhance transparency, accuracy and contextuality of responses. The paper details the end-to-end pipeline, from data collection, preprocessing, to retrieval indexing and response generation, highlighting technical challenges and practical solutions. We aim to offer insights to researchers and practitioners developing similar systems using two distinct approaches: OpenAIs Assistant API with GPT Series and Llamas open-source models . The practical implications of this research lie in enhancing the reliability of generative AI systems in various sectors where domain specific knowledge and real time information retrieval is important. The Python code used in this work is also available at: GitHub.

3626772.3657957.pdf

Evaluating Retrieval Quality in Retrieval-Augmented Generation

Abstract

Evaluating retrieval-augmented generation (RAG) presents chal- lenges, particularly for retrieval models within these systems. Tra- ditional end-to-end evaluation methods are computationally expen- sive. Furthermore, evaluation of the retrieval models performance based on query-document relevance labels shows a small correla- tion with the RAG systems downstream performance. We propose a novel evaluation approach, eRAG, where each document in the retrieval list is individually utilized by the large language model within the RAG system. The output generated for each document is then evaluated based on the downstream task ground truth labels. In this manner, the downstream performance for each document serves as its relevance label. We employ various downstream task metrics to obtain document-level annotations and aggregate them using set-based or ranking metrics. Extensive experiments on a wide range of datasets demonstrate that eRAG achieves a higher correlation with downstream RAG performance compared to base- line methods, with improvements in Kendalls correlation ranging from 0.168 to 0.494. Additionally, eRAG offers significant compu- tational advantages, improving runtime and consuming up to 50 times less GPU memory than end-to-end evaluation.

3644815.3644945.pdf

Seven Failure Points When Engineering a Retrieval Augmented Generation System

Abstract

Software engineers are increasingly adding semantic search capabil- ities to applications using a strategy known as Retrieval Augmented Generation (RAG). A RAG system involves finding documents that semantically match a query and then passing the documents to a large language model (LLM) such as ChatGPT to extract the right answer using an LLM. RAG systems aim to: a) reduce the problem of hallucinated responses from LLMs, b) link sources/references to generated responses, and c) remove the need for annotating documents with meta-data. However, RAG systems suffer from lim- itations inherent to information retrieval systems and from reliance on LLMs. In this paper, we present an experience report on the failure points of RAG systems from three case studies from separate domains: research, education, and biomedical. We share the lessons learned and present 7 failure points to consider when designing a RAG system. The two key takeaways arising from our work are: 1) validation of a RAG system is only feasible during operation, and 2) the robustness of a RAG system evolves rather than designed in at the start. We conclude with a list of potential research directions on RAG systems for the software engineering community.

A_Novel_Linearly_Complex_Extended_Signed_Response- (1).pdf

A Novel Linearly Complex Extended Signed Response-Based Node Authentication Scheme for Internet of Medical Things

Abstract

The Internet of Things (IoT) is considered the future of the Internet due to its immense potential. It has captured the interest of researchers who are exploring new possibilities in this field. Unlike the current internet structure, IoT involves the connection of billions of devices and entails a significant exchange of data, diverse traffic, and resource availability. The increasing use of IoT devices in vulnerable environments has presented two major challenges for researchers: authenticating sensor nodes and ensuring secure data routing. These challenges become more difficult by the presence of wireless sensors in communication devices, where all devices are not authenticated or have up-to-date software. Among the various IoT attacks, the Sybil attack poses a significant threat to the network. Consequently, the conventional solutions used for conventional wireless networks (CWN) do not apply to wireless sensor networks due to the differences in algorithms and associated costs in terms of processing and power consumption. To address these challenges, an extended signed response-based (ESRES) solution is proposed. The proposed framework utilizes a pre-distributed key embedded in a sensor node, while the authentication process employs a dual key-based algorithm with linear complexity. In this mechanism, the node uses pre-distributed keys to respond to a random challenge number sent by the server or sink, thus proving its legitimacy. Since there are different types of IoT networks, such as hierarchical and centralized structures, the proposed authentication scheme is designed to be flexible and

implementable for both types. The performance of the proposed framework is analyzed and evaluated, considering the probability of attack detection with different authentication key pool sizes for the parameters i.e., processing overhead, probability, and power consumption.

A_Novel_Linearly_Complex_Extended_Signed_Response-.pdf

A Novel Linearly Complex Extended Signed Response-Based Node Authentication Scheme for Internet of Medical Things

Abstract

The Internet of Things (IoT) is considered the future of the Internet due to its immense potential. It has captured the interest of researchers who are exploring new possibilities in this field. Unlike the current internet structure, IoT involves the connection of billions of devices and entails a significant exchange of data, diverse traffic, and resource availability. The increasing use of IoT devices in vulnerable environments has presented two major challenges for researchers: authenticating sensor nodes and ensuring secure data routing. These challenges become more difficult by the presence of wireless sensors in communication devices, where all devices are not authenticated or have up-to-date software. Among the various IoT attacks, the Sybil attack poses a significant threat to the network. Consequently, the conventional solutions used for conventional wireless networks (CWN) do not apply to wireless sensor networks due to the differences in algorithms and associated costs in terms of processing and power consumption. To address these challenges, an extended signed response-based (ESRES) solution is proposed. The proposed framework utilizes a pre-distributed key embedded in a sensor node, while the authentication process employs a dual key-based algorithm with linear complexity. In this mechanism, the node uses pre-distributed keys to respond to a random challenge number sent by the server or sink, thus proving its legitimacy. Since there are different types of IoT networks, such as hierarchical and centralized structures, the proposed authentication scheme is designed to be flexible and implementable for both types. The performance of the proposed framework is analyzed and evaluated, considering the probability of attack detection with different authentication key pool sizes for the parameters i.e., processing overhead, probability, and power consumption.

2312.10997v5.pdf

Retrieval-Augmented Generation for Large Language Models: A Survey

Abstract

Large Language Models (LLMs) showcase impres- sive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for

knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain- specific information. RAG synergistically merges LLMs intrin- sic knowledge with the vast, dynamic repositories of external databases. This comprehensive review paper offers a detailed examination of the progression of RAG paradigms, encompassing the Naive RAG, the Advanced RAG, and the Modular RAG. It meticulously scrutinizes the tripartite foundation of RAG frameworks, which includes the retrieval, the generation and the augmentation techniques. The paper highlights the state-of-the- art technologies embedded in each of these critical components, providing a profound understanding of the advancements in RAG systems. Furthermore, this paper introduces up-to-date evalua- tion framework and benchmark. At the end, this article delineates the challenges currently faced and points out prospective avenues for research and development

---

arxiv.pdf

Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability

Abstract

This paper presents an analysis of open-source large language mod- els (LLMs) and their application in Retrieval-Augmented Generation (RAG) tasks, specific for enterprise-specific data sets scraped from their websites. With the increasing reliance on LLMs in natural language processing, it is crucial to evaluate their performance, ac- cessibility, and integration within specific organizational contexts. This study examines various open-source LLMs, explores their in- tegration into RAG frameworks using enterprise-specific data, and assesses the performance of different open-source embeddings in enhancing the retrieval and generation process. Our findings indi- cate that open-source LLMs, combined with effective embedding techniques, can significantly improve the accuracy and efficiency of RAG systems, offering a viable alternative to proprietary solutions for enterprises.