ICRAMCS 2026

THE EIGHTH EDITION OF THE INTERNATIONAL CONFERENCE ON RESEARCH IN APPLIED MATHEMATICS AND COMPUTER SCIENCE
April 23-24-25, 2026 | Marrakech, Morocco

Reliable LLMs for Technical Docs: Modular RAG with Dense Retrieval and Semantic-Aware Chunking

Communication Info

Authors:
Sitayeb SAFOUANE[1]
Aziza El OUAAZIZI[1]
Khalil MAALMI[1]
Faiq GMIRA[1]

[1]Laboratory of Innovative Technologies and Computer Science (LT2I), *Fez, Morocco*

Keywords:
(1) Retrieval-Augmented Generation
(2) Large Language Models
(3) Factual Consistency
(4) Technical Documentation
(5) Vector Databases

Abstract

Large Language Models (LLMs) have achieved significant progress in natural language processing; however, they may generate inaccurate or hallucinated information, particularly in knowledge-intensive domains such as technical documentation support [1]. To address this limitation, this paper proposes DocuBot, a modular framework based on the Retrieval-Augmented Generation (RAG) paradigm that combines external knowledge retrieval with generative models to improve factual reliability [2]. The proposed system follows a two-stage architecture. An offline indexing phase processes technical documents using semantic chunking and stores their embeddings in a vector database. During the online inference phase, the most relevant document segments are retrieved and integrated into the generation process to ground responses in verified information [3].

Experiments conducted on a corpus of car-tracking technical manuals show that the proposed framework improves factual consistency compared with a baseline LLM without retrieval. The results indicate a 31.8% improvement in factual accuracy, reaching 94.1% factual consistency while maintaining acceptable latency.These results demonstrate the effectiveness of RAG-based architectures for improving the reliability of LLM-based assistants in technical documentation and enterprise knowledge management systems [4,5]

References

[1] A. Vaswani et al., "Attention Is All Need," Advances in Neural Information Processing Systems, 2017.

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, 2020.

[3] J. Gao et al., "Retrieval-Augmented Generation: A Survey," arXiv preprint arXiv:2312.10997, 2023.

[4] H. Chen et al., "Agentic RAG: Towards Autonomous Retrieval-Augmented Generation Systems," arXiv preprint arXiv:2402.07015, 2024.

[5] Z. Liu et al., "Advanced RAG: A Survey," arXiv preprint arXiv:2401.07488, 2024.