Leveraging generative AI for knowledge-driven information retrieval in the energy sector

Ali Alsayegh[1*] *and Tariq* Masood[1]

[1]Department of Design Manufacturing and Engineering Management, University of Strathclyde, 75 Montrose Street, Glasgow G1 1XJ, Scotland, UK

Abstract. This paper presents an innovative approach to knowledge management in the energy sector through the development of the Advanced Agent Architecture (AAA). AAA integrates Retrieval-Augmented Generation (RAG) techniques with a tailored local knowledge base (LKM) and web search functionalities, aiming to enhance the accuracy, robustness, and flexibility of information retrieval. We conducted a detailed case study involving a solar power system to evaluate the effectiveness of AAA compared to traditional Large Language Models (LLMs) such as Llama 3. Our results demonstrate that AAA significantly outperforms conventional methods in delivering accurate and relevant answers to complex domain-specific queries. However, the system also shows higher energy consumption and slower response times, identifying critical areas for future research. This study sets the stage for further exploration into optimizing AAA's energy efficiency and processing speed, expanding the range of queries, and providing a more comprehensive benchmarking against traditional systems. Our findings indicate that AAA has the potential to substantially improve knowledge management practices, facilitating more informed decision-making and operational efficiencies in the energy sector.

1 Introduction

This paper investigates the potential of advanced AI techniques to revolutionize knowledge access in the energy sector**.** Traditional information retrieval methods often struggle with the intricate needs of this field, hindering efficient knowledge access for O&M tasks. Generative AI, particularly Large Language Models (LLMs) integrated with Local Knowledge Bases (LKBs), presents a promising solution. However, challenges remain concerning accuracy and LKB limitations. To address these challenges, we introduce the Advanced Agent Architecture (AAA), which merges cutting-edge Retrieval-Augmented Generation (RAG) techniques with LKBs and internet access. This novel approach aims to enhance the accuracy, robustness, and flexibility of Knowledge Management Systems (KMS) in the energy sector. The following sections explore existing KMS limitations, propose a new agent architecture, and demonstrate its effectiveness through a case study.

* Corresponding author: a.alsayegh@strath.ac.uk

2 KMS in energy sectors

The energy sector relies heavily on efficient access to accurate and up-to-date information. To address this need, various KMS have been developed [1]. These systems can take different forms, such as document repositories that store technical specifications, regulations, and best practices. Additionally, expert systems capture and codify the knowledge of human experts, while industry-specific platforms facilitate knowledge sharing among professionals [2]. These KMS offer significant benefits by improving information accessibility and retrieval for energy professionals, enabling faster problem-solving and informed decision-making in areas like renewable energy integration, grid modernization, and energy efficiency [3]. While existing KMS play a valuable role, they often face limitations that hinder their effectiveness[4]. One key limitation is the restricted scope of information. KMS may focus on specific areas within the energy sector, neglecting broader knowledge that could be crucial for understanding complex challenges, such as the integration of renewable energy sources and grid modernization [5]. Furthermore, many KMS operate in isolation, lacking integration with external knowledge sources. This can prevent users from accessing relevant information that resides outside the system's boundaries, hindering knowledge discovery and informed decision-making [6]. Finally, maintaining the accuracy of information within KMS can be challenging, especially as regulations and technologies evolve rapidly in the dynamic energy sector. Generative AI is a subfield of artificial intelligence concerned with creating new data, such as text, code, or images [7]. LLMs are a type of generative AI model trained on massive amounts of text data. These models have shown remarkable capabilities in tasks like text generation, translation, and information summarization [8]. Recent advancements in LLM development, such as GPT-4, Gemini and Llama 3 [9], have yielded models with superior abilities to process information, understand context, and generate human-quality text. Despite their impressive capabilities, LLMs deployed in isolation may not be sufficient for knowledge management in the energy sector [10]. A critical limitation lies in the lack of access to domain-specific knowledge bases. LLMs trained on general text data often struggle with tasks requiring a deep understanding of energy-related concepts, regulations, and technical specifications[11]. This knowledge gap can manifest as inaccurate or misleading responses when confronted with specialised energy sector queries [2]. The overwhelming volume and complexity of information within the energy domain can further exacerbate this challenge, potentially hindering the LLM's ability to discern the most relevant and reliable sources [12]. Furthermore, LLMs are susceptible to a phenomenon known as hallucination, where their outputs contain factual errors not grounded in reality [13]. These errors can stem from patterns or biases within their training data that deviate from factual accuracy [13]. In the context of the energy sector, hallucinations could lead to misguided recommendations or even safety hazards, such as providing incorrect wiring instructions for solar panel installation. However, existing KMS often exhibit limitations like restricted scope and isolation from external knowledge sources, while LLMs, despite their advancements, can be susceptible to factual errors and lack deep understanding of domain-specific knowledge. This necessitates a more robust solution that leverages both domain-specific knowledge and external information access to overcome these challenges.

3 Advanced agent architecture (AAA)

The core of AAA leverages RAG techniques integrated with a local knowledge base (LKM) specifically tailored for the energy sector. Furthermore, AAA has access to external knowledge sources through a web search API. This combined approach aims to enhance the accuracy, robustness, and flexibility of KMS within this domain. Figure 1 illustrates the overall system architecture. The process begins with indexing, where AAA constructs the

LKM using various resources within the energy sector. This LKM serves as a foundation for information retrieval during user interactions. Upon receiving a user query, the LangGraph framework [14] orchestrates a sequence of RAG techniques, including Routing [15], Fallback [16] and Self-correction [17]. Routing directs the query to either the LKM or an external web search depending on the perceived likelihood of finding the answer internally. Fallback refines searches or retrieves additional documents, while Self-correction analyses the generated response to ensure accuracy and eliminate factual errors.

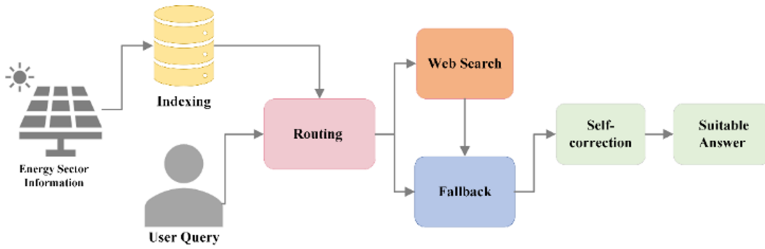

Fig. 1. Proposed Advanced Agent Architecture (AAA**)**

To establish a robust LKM, AAA utilizes a comprehensive indexing approach with several key components. These include a document loader to access and process information from diverse sources, a chunking mechanism to segment large documents into manageable units, a text embedding model to convert textual content into vector representations, and a vector store for efficient retrieval of relevant information based on user queries. AAA leverages Meta Llama 3 [9](released on 18 April 2024), a state-of-the-art LLM with a Mixture of Experts (MoE) architecture. This architecture utilizes multiple specialised sub-models for efficient and accurate language processing. To balance performance and resource efficiency, AAA employs the 8-billion parameter version with 4-bit quantization, enabling faster inference and deployment. Additionally, meticulous prompt engineering ensures effective utilisation of Meta Llama 3 by providing clear and informative prompts within its pre-defined template structure. This maximizes the accuracy, efficiency, and overall effectiveness of AAA's responses. Furthermore, to enhance information access capabilities, AAA integrates the Tavily Search API. This API facilitates web search functionalities by retrieving relevant results from the web. This allows AAA to focus on processing and analysing retrieved information, while Tavily handles web search complexities. Finally, AAA utilizes the LangGraph framework [14] due to its structured approach and predefined control flow. This predefined control flow specifies the tasks the LLM needs to perform at each stage of the interaction, enhancing reliability compared to frameworks placing decision-making burden on the LLM. This structured approach is particularly advantageous for our research considering the utilisation of a smaller, quantized LLM model. LangGraph's well-defined tasks are better suited to smaller LLMs, reducing their overall processing burden, and aligning more effectively with the objectives of our project.

4 Case study

The performance of our AAA was validated through an example case study of a solar power system in a specific energy sector field. A small-scale knowledge base related to solar power systems was created to address this. This knowledge base primarily consisted of Industry-specific knowledge resources such as "Solar Panel Troubleshooting Guide," "Solar Inverter Fault Code Reference Manual," and "Solar Power System Operation and Maintenance Manual". These assets covered the maintenance, operation, and troubleshooting of solar power systems. As this content delves into highly specialised topics, most large language models currently lack the necessary knowledge to handle it effectively. Three representative technical questions related to solar power systems were used to demonstrate the effectiveness of AAA. In addition to evaluating the accuracy and comprehensiveness of the answers provided by AAA compared to a baseline LLM system utilising Llama 3, the study also measured and compared the energy consumption and response time of both systems. This comparison provides valuable insights into the performance characteristics of each approach. Furthermore, the AAA's ability to leverage its web search functionality to address questions beyond the knowledge base was evaluated.

5 Results and discussion

This case study compared the performance of the AAA a baseline LLM system utilising Llama 3 in addressing three representative knowledge-related queries within the domain of solar power systems. These queries covered technical specifications, error code interpretation, and operational decision support. Notably, the operational decision support query required information not readily available in the local knowledge base, necessitating the utilisation of web search capabilities to retrieve the latest weather data. Additionally, the study measured and compared the energy consumption and response time of both systems, providing insights into their relative efficiency.

Table 1. AAA and Llama 3 Q&A on solar power systems related questions with real-time weather data for specific query as of April 28, 2024.

Question 1	What is the maximum AC module ambient temperature for SunPower solar panels?
AAA	*The maximum ambient temperature for SunPower solar panels is 50°C.*
Llama 3	*The maximum ambient temperature (AC) module operating temperature for SunPower solar panels is 85°C (185°F).*
Expected answer	*Max. AC Module Ambient Temp. +50°C*
Question 2	What is error code 053?
AAA	*The error code 053 indicates that the slave DSP within the inverter system has detected a grid voltage exceeding the inverter allowable upper limit*
Llama 3	*Error code 053 typically indicates a file system error or a problem accessing resources on a system.*
Expected answer	*The slave DSP detects that the grid voltage exceeds inverter allowable upper limit*
Question 3	I'm planning to do some maintenance work on some solar panels today at the shotwick solar park, uk. Based on the current weather forecast for the area, will there be any rain or strong winds that could impact the cleaning process ?
AAA	*Based on the current weather forecast, there is a slight chance of rain with 0.13mm of precipitation expected, but no strong winds are predicted. This should not significantly impact your cleaning process at the Shotwick Solar Park in the UK.*
Llama 3	*According to the current weather forecast, Shotwick Solar Park in Northern Ireland is expected to have overcast conditions with a slight chance of light showers and moderate winds (15-20 mph).*
Expected answer	*no wind, slight chance of minimal rain, cleaning should proceed without significant impact.*

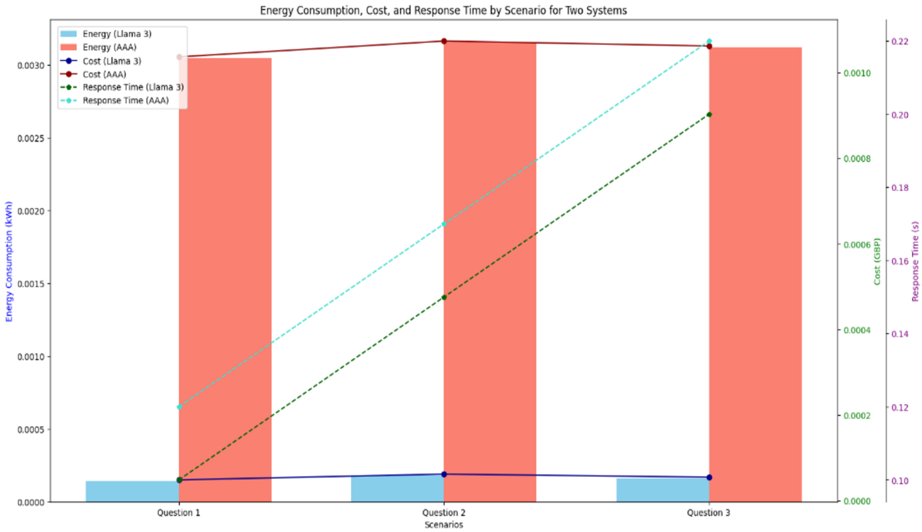

Fig. 2. Energy Consumption, Cost, and Response Time by Scenario for Two Systems

The case study findings demonstrate that the AAA system, while exhibiting higher energy consumption and slower response times as shown in the Figure 2, consistently delivered accurate and relevant answers to all three solar power-related questions as demonstrated in Table 1, whereas the LLM system (Llama 3) provided inaccurate or incomplete information in all three cases. This stark contrast highlights the AAA system's ability to effectively leverage web search functionalities, when necessary, a capability that proved particularly advantageous in the third question where the required information was not readily available within its local knowledge base. By seamlessly integrating web search, the AAA system accessed real-time weather data, enabling it to provide a comprehensive and accurate response. However, the increased operational costs and energy usage associated with the AAA system, compared to the LLM system's lower consumption and faster responses, highlight areas for improvement. Given the AAA's unique strength in effectively integrating external data sources through web search and leveraging its specialised LKB, future research should prioritize optimizing its energy efficiency and processing speed to enhance its real-world applicability within the energy sector. Expanding the research scope to encompass a broader range of solar power-related queries and incorporating more comprehensive benchmarking of both systems' performance will provide deeper insights into their relative efficiencies and guide further refinement of the AAA system for practical use.

6 Conclusion

This study presents a comprehensive evaluation of the AAA and its potential to revolutionize knowledge management within the energy sector. The AAA leverages RAG techniques, combining a specialized LKB tailored to the domain with the dynamic integration of web search capabilities. Notably, this paper marks the first documented use of Meta Llama 3 as a RAG agent, signifying a significant advancement in applying cutting-edge AI within this field. While the case study using a solar power system demonstrates the AAA's enhanced accuracy and robustness in handling domain-specific queries, it also reveals higher energy consumption and slower response times compared to traditional systems. These findings highlight crucial areas for further research, particularly focusing on optimizing the system's energy efficiency and processing speed to enhance its real-world feasibility. Future studies should expand the range of tested queries and incorporate detailed comparative analysis with traditional systems. This comprehensive approach will thoroughly assess and refine the AAA's performance, ensuring its seamless integration into existing knowledge management frameworks and its potential to significantly boost decision-making and operational efficiency in the energy sector.

References

1. L. Jie, L. Gang, M. Xin, *The Construction of Knowledge Management System and Model in Electric Power Enterprises*. In Proceedings of the 2010 International Conference of Information Science and Management Engineering, 7-8 Aug. 2010, pp. 79-84 (2010)
2. K. J. Kumar, R. Sharma, J. Inf. Knowl. Manag **21 (03)**, 2250040 (2022)
3. E. G. Ochieng, O. O. Ovbagbedia, T. Zuofa, R. Abdulai, W. Matipa, X. Ruan, A. Oledinma, Inf. Technol. People, **31** (1), 527-556 (2018)
4. T. Masood, J. A. Erkoyuncu, R. Roy, A. Harrison, CIRP J. Manuf. Sci. Technol, **7** (2), 83-96 (2014)
5. C. -P. Simion, C. -A. Verdeș, A. -A. Mironescu, F. -G. Anghel, Energies, **16** (4), 1960 (2023)
6. A. V. Chernov, V. A. Chernova, T. V. Komarova, *The Usage of Artificial Intelligence in Strategic Decision Making in Terms of Fourth Industrial Revolution.* In Proceedings of the 1st International Conference on Emerging Trends and Challenges in the Management Theory and Practice (ETCMTP 2019), 22-25, (2020)
7. S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative AI. Bus. Inf. Syst. Eng, **66**, 111-126 (2023).
8. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *Language models are few-shot learners.* Adv. Neural Inf. Process. Syst, **33**, 1877-1901 (2020)
9. Meta. Llama 3. Available online: https://llama.meta.com/llama3/ (accessed on 24/4).
10. L. Dong, S. Majumder, F. Doudi, Y. Cai, C. Tian, D. Kalathi, K. Ding, A. A. Thatte, L. Xie, Joule **8** (6), 1544-1549 (2024)
11. H. Wang, Y. F. Li, *Large Language Model Empowered by Domain-Specific Knowledge Base for Industrial Equipment Operation and Maintenance.* In Proceedings of the 2023 5th International Conference on System Reliability and Safety Engineering (SRSE), 20-23 Oct. 2023, pp. 474-479 (2023)
12. F. A.AlSelami, I. M.ELEmary, H. M.Alamoudi, Int. J. Eng. Res. Technol **13** (4), 744 (2020).
13. Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, P. Fung, *Towards Mitigating LLM Hallucination via Self Reflection.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (2023)
14. Langchain. LangGraph. Available online: https://python.langchain.com/docs/langgraph/ (accessed on 25/4).
15. S. Jeong, J. Baek, S. Cho, S. J. Hwang, J. C. Park, *Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity.* in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2024m Mexico City, Mexico 7036-7050 (2024)
16. S. -Q. Yan, J. -C. Gu, Y. Zhu, Z. -H. Ling, *Corrective Retrieval Augmented Generation.* arXiv preprint arXiv:2401.15884 (2024)
17. A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, *Self-rag: Learning to retrieve, generate, and critique through self-reflection.* arXiv preprint arXiv:2310.11511 (2023)