Wikatoni : An Agentic AI System for Energy Engineering Workflows

Sampath Rajapaksha[1,2], **Nirmalie Wiratunga**[1], **Ikechukwu Nkisi-Orji**[1], **Tim Clarke**[2], **Fraser Kerr**[2]

[1]Robert Gordon University, Aberdeen, UK
[2]Katoni Engineering, Aberdeen, UK

Abstract

Capturing expertise and enabling efficient information retrieval are critical in the energy sector, where high staff turnover can lead to significant knowledge loss. Retrieval Augmented Generation (RAG) offers a solution by grounding Large Language Model (LLM) outputs in documented sources, but its effectiveness is limited by reliance on general-purpose embeddings. We present Wikatoni, an agentic AI system for energy engineering workflows that integrates a novel domain-specific embedding model. Wikatoni combines fine-tuned embeddings with agentic RAG, metadata filtering, and hybrid retrieval to improve document search, automated reporting, and workflow efficiency. Evaluation on internal enterprise offshore energy data shows that the domain-adapted embedding improves recall by 10%, and Wikatoni agentic RAG further increases answer accuracy by 14% compared to vanilla RAG with the base embedding model, achieving the best overall performance in context recall, faithfulness, and answer accuracy.

Model — https://hf.co/Sampath1987/EnergyEmbed-v1

Datasets —
https://hf.co/collections/Sampath1987/energy-datasets

Introduction

A major challenge in the energy sector is the loss of critical knowledge due to staff turnover, which can lead to revenue decline, reduced productivity, workflow disruptions, and missed business opportunities (Sumbal et al. 2018). Knowledge, including scientific, technological, and managerial types, is essential for strategy and operations (Edwards 2008), and effective retention requires identifying critical knowledge, transferring undocumented essentials, and integrating retained knowledge into business processes (Edwards 2008). Documentation and standardisation are crucial for capturing knowledge and enabling efficient information retrieval.

Retrieval-Augmented Generation (RAG) addresses limitations of Large Language Models (LLMs) in organisation-specific contexts by grounding responses in external knowledge bases (Lewis et al. 2020; Rajapaksha, Rani, and Karafili 2024). A key component is the embedding model, which retrieves relevant documents. General-purpose embeddings often underperform in specialised domains due to unique vocabulary and complex semantics (Shiraee Kasmaee 2025), while domain-specific embeddings improve semantic search and question answering (Tang and Yang 2024). Previous work introduced an energy-sector embedding model but did not release the model or training data (Haddadian, Chen, and Shor 2025).

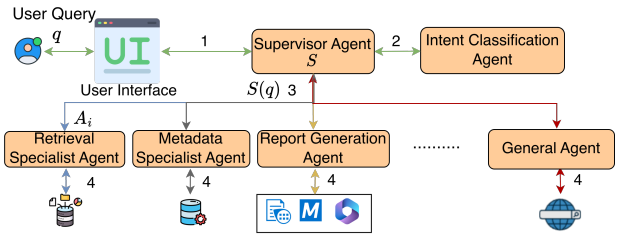

Figure 1: Wikatoni System Architecture

To overcome these limitations, we present Wikatoni, an agentic AI system for energy engineering workflows. It incorporates a novel embedding model fine-tuned on a domain-specific dataset, both openly released. By combining metadata filtering, hybrid retrieval, and reranking, Wikatoni enhances document search, processing, and automated reporting for complex engineering tasks in the energy industry.

Embedding Model Fine-tunning

To fine-tune the embedding model, we built a domain-specific dataset by collecting 15,000 abstracts from oil and gas conference papers (Society of Petroleum Engineers 2025), technical publications (International Association of Oil & Gas Producers 2025), OEUK reports (Offshore Energies UK 2025), and the Volve field dataset (Equinor 2025). Anchor, positive, and negative triplets were generated for contrastive learning, with semantic document chunks as positives and up to five questions per chunk as anchors using OpenAI 4o-mini to capture domain-specific vocabulary and semantics. Negatives were selected via hard-negative mining (de Souza P. Moreira et al. 2025). After comparing embedding models using cosine accuracy and token limits, gte-multilingual-base (Zhang et al. 2024) was chosen as the base

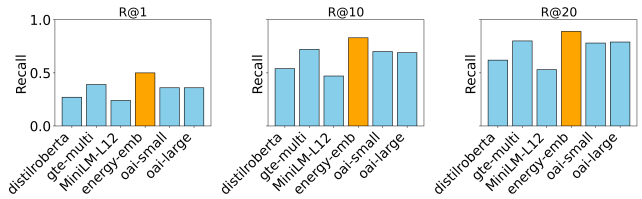

Figure 2: Recall@K Comparison Across Models

Model	CR	F	AA
RAG-distilroberta	0.73	0.90	0.57
RAG-MiniLM-L12	0.65	0.63	0.46
RAG-oai-small	0.86	0.93	0.69
RAG-oai-large	0.88	0.93	0.71
RAG-gte-multi (base model)	0.85	0.95	0.69
Wikatoni-RAG (energy_emb)	0.92	0.96	0.77
Wikatoni-Agentic (energy_emb)	**0.94**	**0.97**	**0.83**

Table 1: Evaluation results for Wikatoni

model. MultipleNegativesRankingLoss objective was used for training over one epoch, as longer runs consistently led to overfitting and reduced retrieval generalization.

System Overview

The Wikatoni architecture integrates embedding techniques, advanced retrieval strategies, and ReAct agents (Yao et al. 2023) to enhance reliability, relevance, and reduce hallucinations. The overall system design, illustrated in Figure 1, supports a range of practical use cases. This utilises the multi-agent supervisor architecture, where specialised agents are coordinated by a central supervisor agent. The supervisor agent controls all communication flow and task delegation, making decisions about which agent to invoke based on the current context and task requirements.

Supervisor Agent The supervisor agent S receives a user query q and analyses it to route the request to the appropriate sub-agent. Instead of deciding routing directly, S leverages the intent classification agent, which provides the query intent, candidate agents, metadata, and confidence score. Based on this output, S either routes the query to the best-matched sub-agent A_i ($S(q) \rightarrow A_i$) or requests follow-up information if the query is ambiguous. This process is illustrated in Steps 1 and 2 of Figure 1. The system includes ten sub-agents responsible for tasks such as internal data retrieval, metadata retrieval, AI report automation, Microsoft Business Central data analysis, employee expertise analysis, and handling general queries.

Retrieval Agents The retrieval specialist and metadata specialist are core modules designed to extract information from organisation-specific documents. Both agents use a retrieval tool connected to vector databases containing a wide variety of engineering documents. The fine-tuned oil and gas embedding model is employed for database creation and information retrieval. The retrieval specialist focuses on factual queries using similarity-based retrieval with agentic RAG using ReAct with advanced techniques such as hybrid retrieval and reranking, while the metadata specialist performs broader analyses across multiple documents. For this purpose, the metadata specialist processes entire document sets matching defined metadata and leverages long-context (LC) LLMs with ReAct.

Other Sub-agents Additional sub-agents address domain-specific workflows. For example, report generation in energy engineering often requires extracting data from complex CAD drawings. Wikatoni fine-tunes OCR and vision models (Mittal and Garg 2020; Jung, Kim, and Jain 2004) to extract relevant information, which is then combined with chain-of-thought reasoning to generate accurate reports. The general agent handles miscellaneous queries, retrieving online data when necessary. All sub-agents are carefully designed to meet the diverse requirements of the oil and gas domain. For all tasks, Wikatoni employs different closed-source LLMs as generator models, selected based on the specific use case. These models are hosted on Microsoft Azure to ensure data privacy and security.

Evaluation

To evaluate the embedding model, a test set was created using internal enterprise offshore energy data, distinct from the publicly available datasets used for training. The fine-tuned energy-emb was compared against the base model of gte-multilingual-base (gte-multi), all-distilroberta-v1 (distilroberta), all-MiniLM-L12-v2 (MiniLM-L12), and OpenAI's small and large embedding models (oai-small and oai-large). Figure 2 shows the Recall@K comparison, where energy-emb consistently outperforms all models, exceeding the base and OpenAI large models by 10% at K=20.

To assess the accuracy of the agentic AI framework Wikatoni, 300 real user queries with expert-annotated ground truth were evaluated using the RAGAS framework (Es et al. 2024) with K=20 for Context Recall (CR), Faithfulness (F), and Answer Accuracy (AA)(Yu et al. 2024). As shown in Table 1, vanilla RAG performance varies significantly across embedding models. Integrating energy_emb into vanilla RAG (Wikatoni-RAG) substantially improves all metrics, while Wikatoni-Agentic, which combines energy_emb with agentic RAG, achieves the best results overall. This demonstrates that domain-specific embeddings and agentic reasoning together deliver superior performance in the oil and gas domain.

Conclusion

Wikatoni effectively addresses core limitations of existing RAG frameworks in enterprise energy settings by introducing domain-specific embedding model and using agentic AI system using ReAct to improve both retrieval and generation quality. Wikatoni not only enhances transparency and trust but also significantly boosts the efficiency of knowledge-intensive energy engineering workflows.

Acknowledgments

This research was conducted as part of a Knowledge Transfer Partnership (KTP) between Robert Gordon University and Katoni Engineering Ltd. The authors gratefully acknowledge funding support from Innovate UK under Grant KTP Reference Number KTP13834.

References

de Souza P. Moreira, G.; Osmulski, R.; Xu, M.; Ak, R.; Schifferer, B.; and Oldridge, E. 2025. NV-Retriever: Improving text embedding models with effective hard-negative mining. arXiv:2407.15831.

Edwards, J. S. 2008. Knowledge management in the energy sector: review and future directions. *International Journal of Energy Sector Management*, 2(2): 197–217.

Equinor. 2025. Volve Field Data Set. https://www.equinor.com/energy/volve-data-sharing.

Es, S.; James, J.; Anke, L. E.; and Schockaert, S. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.

Haddadian, K.; Chen, S.; and Shor, R. 2025. Enhancing Question-Answering in Energy Engineering Via a Domain-Specific Embedding Model in an Open-Source Rag Pipeline: A Case Study on Geothermal Energy. *Available at SSRN 5357219*.

International Association of Oil & Gas Producers. 2025. IOGP Bookstore. https://www.iogp.org/bookstore/. Accessed: 2025-09-15.

Jung, K.; Kim, K. I.; and Jain, A. K. 2004. Text information extraction in images and video: a survey. *Pattern recognition*, 37(5): 977–997.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Mittal, R.; and Garg, A. 2020. Text extraction using OCR: a systematic review. In *2020 second international conference on inventive research in computing applications (ICIRCA)*, 357–362. IEEE.

Offshore Energies UK. 2025. OEUK Reports. https://oeuk.org.uk/product-category/oeuk-reports/. Accessed: 2025-09-15.

Rajapaksha, S.; Rani, R.; and Karafili, E. 2024. A RAG-based question-answering solution for cyber-attack investigation and attribution. In *European Symposium on Research in Computer Security*, 238–256. Springer.

Shiraee Kasmaee, A. 2025. *Domain-Specific Text Embedding Models for Information Retrieval*. Ph.D. thesis, McMaster University.

Society of Petroleum Engineers. 2025. SPE Conferences. https://onepetro.org/SPE/pages/conferences.

Sumbal, M. S.; Tsui, E.; Cheong, R.; and See-to, E. W. 2018. Critical areas of knowledge loss when employees leave in the oil and gas industry. *Journal of Knowledge Management*, 22(7): 1573–1590.

Tang, Y.; and Yang, Y. 2024. Do we need domain-specific embedding models? An empirical investigation. *arXiv preprint arXiv:2409.18511*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629.

Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; and Liu, Z. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, 102–120. Springer.

Zhang, X.; Zhang, Y.; Long, D.; Xie, W.; Dai, Z.; Tang, J.; Lin, H.; Yang, B.; Xie, P.; Huang, F.; et al. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1393–1412.