

Empowering Self-Powered Sensors via Agentic RAG: A Domain-Specific LLM Framework for Building Physics Retrieval

(Sobhan Saadatnasab)¹,

¹ (Sobhan Saadatnasab), Email: (ssaadatnasab@connect.ust.hk)

ABSTRACT

Wireless sensor networks deployed in built environments face a critical bottleneck: reliable in-situ powering. Energy harvesting, whether from thermal gradients, ambient airflow, or mechanical vibrations—demands precise, context-aware calculations of building physics parameters such as heat flux, air velocity, temperature differentials, and fluid properties. However, general-purpose Large Language Models lack the domain-specific reasoning required to retrieve and compute these engineering metrics from technical literature, codes, and equipment schedules, as they consistently fail to accurately interpret structured tables, complex mathematical formulas, and chart-based empirical data.

To bridge this gap, we present a domain-specific LLM framework powered by an advanced multimodal Agentic Retrieval-Augmented Generation (RAG) pipeline. Our framework implements an end-to-end pipeline that converts heterogeneous technical literature into a high-quality Markdown corpus and leverages vision-language extraction to index information from figures, schematics, and images.

A core innovation lies in its agentic architecture, which features autonomous planning and a dedicated pathway that executes Python code within a secure sandbox. This ensures strict mathematical precision for engineering calculations—a necessity for tasks such as power density estimation, conversion efficiency analysis, and storage sizing in energy-constrained sensors.

We rigorously validate our approach using a comprehensive benchmark of 1,487 technical questions spanning 12 core building science domains (heat transfer, HVAC, thermodynamics, fluid mechanics, and automation) across five question types and three difficulty levels. Our evaluation of 15 state-of-the-art LLMs reveals a base accuracy of only 58.7%, confirming the inadequacy of off-the-shelf models. To address this, we implement a hybrid retrieval stack combining BM25 sparse retrieval, dense MiniLM embeddings (fused at 0.4/0.6), MMR diversification, and a bge-reranker-v2-m3 cross-encoder. Extending this with an Agentic RAG workflow yields a +4% overall accuracy boost and a remarkable +8.2% improvement on hard questions, with specific models like Claude Opus 4.6 gaining up to +22.5 points over their base versions.

By integrating these capabilities, our approach delivers a robust, high-fidelity reasoning and retrieval tool specifically designed for specialized engineering domains. This framework empowers engineers to rapidly query, calculate, and verify energy-harvesting potential from building data, significantly accelerating the design and optimization of self-powered, maintenance-free wireless sensing systems.