A RULE-BASED AI CHATBOT FOR HEN DESIGN WITH A MINIMAL DATABASE AND AN EXTENDED KNOWLEDGE BASE

Roberto Melli[1]*, Enrico Sciubba[2]

[1]MPES s.r.l., Rome, Italy

[2]Niccolò Cusano University, Rome, Italy, enrico.sciubba@fondazione.uniroma1.it

*Corresponding Author: xpertsystems@mpessrl.it

ABSTRACT

Four decades ago, Gaggioli, Sama, and Qian published a series of 10 "Second Law Guidelines for design and process engineers". Sama had previously (1980-1983) published part of this material as a compilation of "second law errors" to avoid in the design of energy conversion systems. The list was rearranged several times until a revised version containing 21 rules was published by Sama and Szargut in 1995. This final version came to be known as "the Sama-Szargut rules", and consists in a series of well-formulated and insightful suggestions that reflect a thermodynamicist's view about "best design practices": the subsumed principle being that the -technically and economically- most convenient process structure is one that minimizes the overall irreversibility within the limits imposed by a set of technological constraints. Notably, the concept of "optimal system" is completely absent from the emphasis. The difference between the Sama-Szargut rules and the usual engineering design manuals is the extensive inclusion of second law reasoning into the former.

Focus of our investigation is indeed the conceptual design of Heat Exchanger Networks (HEN): considering that it is a quite mature field with an extensive database of empirical results and a set of well-established design procedures, it is tempting to seek advice from one of the so-called "AI assistants", be it generalists or specialized: this seems to be indeed the direction industry is currently taking. As an alternative, we suggest instead that the Sama-Szargut rules be adopted as the Knowledge Base for an Inference Engine that makes little or no use of Data Mining and Database Scanning but works instead on a higher, "conceptual" basis: once the design specifications are assigned, the system "applies" the rules systematically and uses them to construct a Knowledge Graph that can be inspected by a human operator/designer.

The procedure is conceptually straightforward (but requires accurate programming!): the Sama-Szargut rules are subjected to the Graph Retrieval-Augmented Generation (Graph-RAG) methodology that integrates a knowledge graph with an LLM reasoning layer. This type of approach is particularly advantageous for Second Law-based design, where constraints, thermodynamic principles, and structural relationships interact in non-linear ways. In fact, a primary advantage of Graph-RAG is its capacity to encode thermodynamic knowledge within structured relational frameworks.

Some examples of applications to realistic design cases are discussed.

1 INTRODUCTION

This paper proposes the implementation of the Graph Retrieval-Augmented Generation (Graph-RAG) methodology to analyze the Sama–Szargut Second-Law-based criteria specifically applied in the design of heat exchanger networks (HENs) [1,2]. Such an application of Graph-RAG is a novel idea, but it can be rigorously justified on methodological, thermodynamic, and engineering grounds. Creating a correspondence between Second Law-based rules and graph-structured representations of engineering knowledge improves accuracy in design optimization, ensures abidance by thermodynamic constraints, and enhances the overall quality of reasoning [3].

Conducting a qualitative analysis before performing quantitative engineering calculations clarifies the nature of the technical and economic problems before any computations are performed. This is critical for developing an in-depth understanding of the underlying phenomena, identifying key variables, and formulating appropriate hypotheses, thereby improving accuracy, precision, data interpretability, and overall process best practice.

2 HOW GRAPH-RAG METHODOLOGIES FACILITATE THE IMPLEMENTATION OF THE SAMA-SZARGUT SECOND LAW RULES

The Sama–Szargut principles are based on exergy efficiency considerations and do not rely on simple numerical formulas. Instead, they consist of qualitative and logical guidelines developed from the Second Law of Thermodynamics. Performing a qualitative analysis helps clarify the nature of the engineering problem before any calculations are undertaken, thereby narrowing the scope of possible solutions. This approach identifies objectives, constraints, and (physical and logical) boundaries to ensure that the problem is completely defined and well-posed, so that efforts are directed toward solving an appropriately framed challenge, rather than optimizing an ill-posed or incomplete problem.

Engineering systems typically involve multiple interacting physical, chemical, economic, and operational factors. Qualitative reasoning assists in distinguishing between dominant and negligible mechanisms, thereby preventing unnecessary complexity in modeling and avoiding the application of detailed numerical methods to physically irrelevant aspects. This procedure, called "preliminary identification of non-relevant variables", lowers the dimensions of the solution space therefore reducing computational requirements and enhancing the interpretability of quantitative results.

Applying core engineering principles to the formulation and implementation of Second-Law-based rules—such as avoiding high-temperature heat rejection or the use of high-quality heat for low-grade processes—requires careful consideration of the specific context. The correct application of these rules depends on various factors, including process topology, stream roles (hot/cold, utility or process streams), temperature ranges, overlaps, as well as economic and operational constraints. Contextual engineering explicitly structures this information to ensure that rules are applied conditionally, rather than purely mechanically.

3 THE DOMAIN CONTEXT ENGINEERING METHODOLOGY APPLIED TO SAMA-SZARGUT SECOND-LAW RULES FOR HEAT EXCHANGER NETWORKS SYNTHESIS AND ANALYSIS

Context-based engineering[4,5] can be systematically applied to the Sama–Szargut second-law–based principles. We will analyze this approach, with a particular focus on transitioning from rule-based methods to context-aware strategies in the synthesis and optimization of Heat Exchanger Networks (HEN).

The Sama–Szargut principles are localized heuristics based on Second Law considerations and exergy efficiency, emphasizing practices such as minimizing high-temperature heat rejection and avoiding the use of high-quality heat for low-grade applications. However, their effective implementation is highly context-specific: it can be decomposed into several layers and depends on factors such as process configuration, the role of different streams (e.g., hot or cold, utility or process), temperature ranges and overlaps, as well as economic and operational constraints.

For example, for HENs, the layered structure [6] can be organized as follows:

- Thermodynamic Context, like "*do not discard heat at high temperatures to the ambient or to cooling water, and neither heat refrigerated streams with hot streams or with cooling water; try to match streams where the final temperature of one is close to the initial temperature of the other; any counter-current process is generally more thermodynamically efficient than its parallel equivalent*."
- Structural Context, like "*locating hot process streams, cooling and hot utilities, refrigerated streams, phase changes, and topology constraints*."

- Functional Context, like "*consider batch vs. continuous operation; any counter-current process is generally more thermodynamically efficient than its parallel equivalent; minimize the throttling of steam or other gases; eliminate leaks in pipelines, valves, and combustion chambers*."[1]

As a result, the application of Context Engineering transforms the Sama–Szargut rules from static thermodynamic assertions into dynamic, context-related engineering constraints. It organizes streams by energy quality, highlighting critical contexts.

4 PRACTICAL IMPLEMENTATION OF CONTEXT ENGINEERING IN HEN DESIGN

Large language models (LLMs) operate probabilistically and lack domain knowledge. RAG mitigates this limitation by:

- Performing a semantic search over structured data
- Retrieving high-relevance documents, or graph nodes
- Supplying evidence-grounded content to the model

This enables the model to transition from a probabilistic approach to a context-dependent approach.

Retrieval-Augmented Generation (RAG) [7,8,9,10,11] can be effectively situated within a context engineering framework as a structured mechanism for acquiring, validating, and synthesizing external knowledge before and during the generative process. It does so by incorporating information retrieval processes directly into the generative architecture of large language models, integrating their output with contextually relevant external data by leveraging high-dimensional vector representations. This integration facilitates the generation of responses from verifiable, external knowledge databases, including dynamically changing data sources, effectively extending the model's capability from purely generative inference to context-aware response synthesis.

RAG is currently used to quickly access technical documentation, standards, and historical maintenance data for machinery diagnosis, design automation, or troubleshooting complex mechanical systems, and transform the data into high-dimensional vector representations, index these vectors within a similarity search structure, and employ the model to perform approximate nearest neighbor retrieval of analogous vector segments during inference. Consequently, typical applications of RAG include, among others, responding to user inquiries in both open and restricted domains, improving the accuracy and relevance of chatbots, and summarizing documents by using retrieved data as contextual information.

Although RAG systems are highly effective, they have certain limitations. While embedding representations effectively captures semantics, they lack inherent structural or relational knowledge. Specifically, these embeddings do not encode explicit domain-specific constraints or logical relationships. Additionally, a significant limitation is their limited ability to perform comprehensive reasoning across multiple documents.

Graph-RAG combines a knowledge graph with an LLM reasoning layer. This integration is particularly beneficial for Second-Law-based design, where constraints, thermodynamic principles, and structural relationships interact in complex, non-linear ways. A key benefit of Graph-RAG is its ability to effectively incorporate thermodynamic knowledge within structured, relational frameworks.

A knowledge graph can encode these relationships explicitly:

- Nodes: streams, utilities, exchangers, temperature levels, etc.
- Edges: "supplies exergy to", "receives heat from", "violates rule X", "eligible match", etc.

This process encodes implicit domain-specific expertise into a formal, machine-readable representational format (chunks). By their very nature, the Sama–Szargut Second Law rules operate on relational, topological, and qualitative properties of energy systems, and their direct translation into quantitative design is conceptually complicated, computationally laborious and requires additional

---

[1] Notice the repetition of the "counterflow" argument in Thermodynamic and Functional contexts: this is neither casual not introduces real redudancy, because the"reasoning activities" in each context proceed along parallel paths until the final synthesis.

case-specific instructions (e.g.:"consider a gas-turbine based cogenerating plant"; or "Consider a Heat Exchangers Network design problem", etc.) [19]. Graph-RAG is well-suited to represent, retrieve, and reason over such properties in an explicit, scalable, and physically meaningful way.

5 GRAPH-RAG IMPLEMENTATION AND TEXT CHUNK SIZE OPTIMIZATION WORKFLOWS

5.1 Description of the Problem

Graph RAG is an advanced retrieval-augmented generation methodology that constructs knowledge graphs through the association of text segments via high-dimensional vector similarity metrics. This approach facilitates the encoding of semantic relationships and leads to a more comprehensive and contextual understanding, while offering more accurate and detailed responses compared to traditional RAG systems that may semantically identify relevant document segments but often overlook important interrelated information necessary for constructing comprehensive responses to user queries.

In RAG, documents are subdivided into batches ("segments"), each of which is independently embedded as a high-dimensional vector within the "relevant data" space. As a result, the original structural elements—such as headings, narrative flow, and references between paragraphs—are no longer preserved.

Graph-based RAG addresses this limitation by explicitly modeling the relational structure among information chunks within the document. Rather than merely retrieving similar content, it constructs a knowledge graph that maps out the connections between all relevant data points, enabling an integrated understanding of the document's contextual relationships for enhanced answer generation.

The Graph RAG Vector similarity system workflow is presented in. Figure 1.

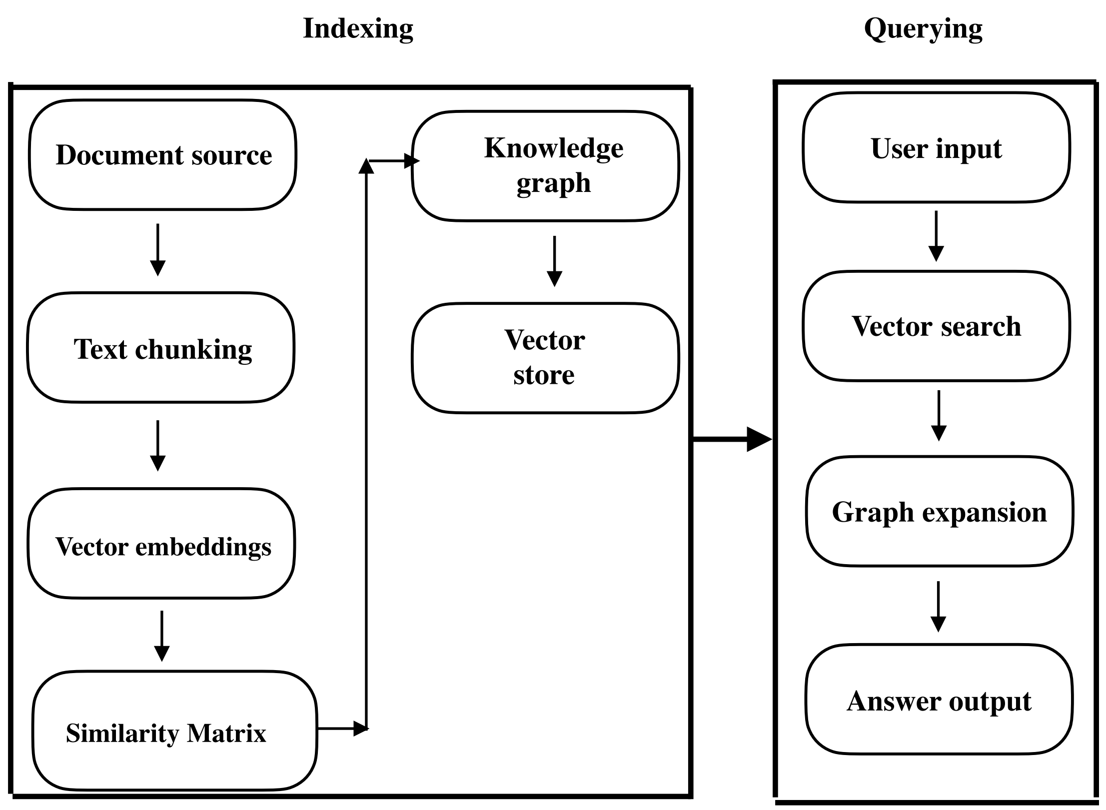

Figure 1: Graph RAG Vector similarity system workflow

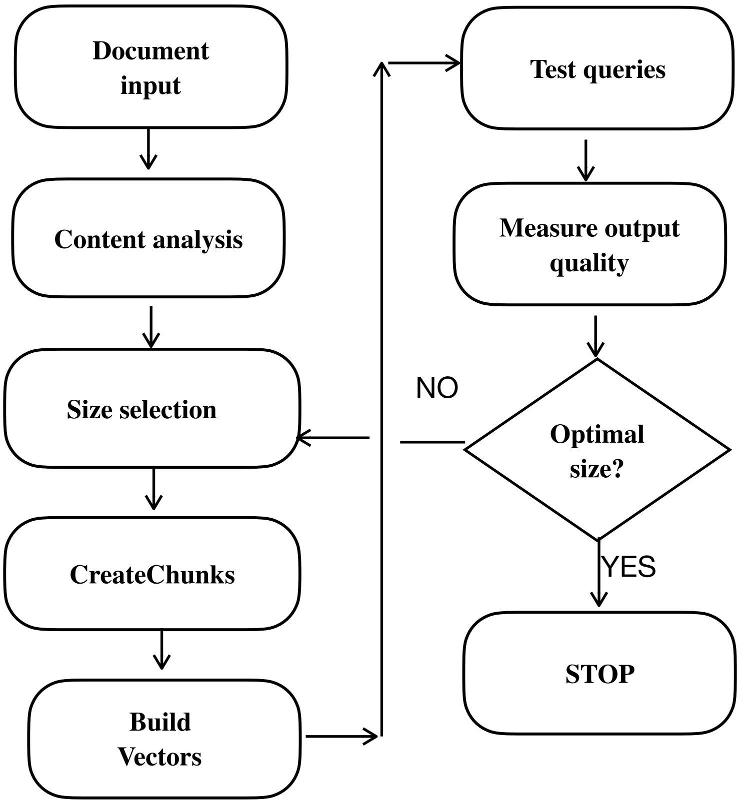

Figure 2: Chunk size selection workflow

5.2 RAG Text Chunk Size Optimization

The most critical factor in the RAG indexing process is the optimal choice of the text chunk size. The key to an optimal selection lies in balancing retrieval accuracy with comprehensive context retention, tailored to the specific application scenario. An excessively narrow chunk size may lead to incomplete context capture, thereby compromising the relevance of retrieved information, whereas a too broad chunk size can result in overly generalized responses lacking specificity. So, the first step in the implementation of RAG is the choice of an optimal Test Chunk size. The workflow to find the possible best Chunk size is presented in Figure 2: it resembles the training scheme of an ANN.

6 DOCUMENT PROCESSING AND PDF EXTRACTION

Graph RAG operates on the documents partitions it extracted from the textual Database (the above mentioned "chunks") and transforms them into vector embeddings by establishing connections between semantically similar segments that meet and exceed a defined similarity threshold. This process effectively creates a structured, semantic representation of the entire document collection.
The resulting graph structure enables cross-document information synthesis and relational analysis across multiple documents and facilitates the understanding of relationships between entities, thereby facilitating complex query resolution through multidimensional connectivity and contextual reasoning within the assembled knowledge graph.
The set up of a RAG workflow as represented in Figure 1 starts with the selection of Chunk size, that is the parameter determining the fragment granularity used for textual segmentation in RAG systems. An excessively granular (small) chunking approach (e.g. under 128 tokens) may omit contextual coherence, thereby missing overall comprehension and retrieval accuracy. Conversely, overly coarse segmentation by specifying large chunks (e.g. over 1024 tokens) can obscure pertinent details and reduce specificity, hindering precise information retrieval. Optimal calibration of chunk size enhances the model's ability to accurately understand and retrieve relevant content, thereby significantly improving system performance. In this study a chunk size of 256 tokens has been selected, as common practices have been proven to be an effective average value. Although it is possible that applying the selection procedure sketched in Figure 2 to a larger -or more specific- database, a

different "optimal" chunk size may emerge, the a priori choice made here suffices for the demonstration of the method.

Another important parameter is the so-called "Cosine Similarity". It is a metric that quantifies the cosine of the angle between two non-zero vectors within a high-dimensional vector space. It is commonly employed to evaluate the degree of similarity between vectors extracted from textual documents, serving as a better indicator of the relative importance of a term within a specific document. It represents the translation of a conceptual similarity into a much easily understandable (and intuitively quantifiable) Euclidean similarity that compares the distance between two points in a N-dimensional space with the "degree of parallelism" of their respective vectors.
Batch size refers to the number of examples processed simultaneously during training. An appropriate batch size is essential for balancing memory usage and training efficiency. A common rule of thumb is to start with a batch size between 32 and 128.
The methodology is illustrated using Reference 6, specifically the first seven pages because they focus on the Sama-Szargut rules. The constraint of a seven-page bound is due to the computational capacity constraints in the concurrent processing of multiple batches.

Table 1: Graph RAG quantitavive key values (* pre-assigned values)

Chunk size	256* tokens
Chunk overlapping	30* / 50* tokens
Split 7 pages into	94 chunks
Average chunk size	199 characters
Batch size	32* tokens
Embeddings	94
Dimensions of each embedding	1536
Similarity Matrix shape	(94, 94)
Similarity values range	From 0.673 to 1.000
Nodes added to the graph	94
Total pairs	4371
Edges create from 4371 possible pairs	1140
Graph density	0.261
Average similarity of connected chunks	0.822

The most interconnected segments typically represent the key concepts, as they tend to appear in contexts similar to numerous other segments (they are, so to say, "more popular" throughout the Database).

Here is the list of the most connected chunks (as the beginning of the sentence found in the document):
Most Connected Chunks:
- Node 26 (60 connections): with the T). The rule applies with obvious modifications to chemical, electrical and fluid processes …

- Node 44 (47 connections): method) has convincingly shown that the separate optimization of one unit of an energy system does …
- Node 13 (46 connections): referred to as “the Sama-Szargut rules”. The rules reflect a thermodynamicist’s idea that the “best d…
- Node 40 (46 connections): the chain of processes and therefore ensures a reduction of the exergy destruction within the entir…
- Node 68 (46 connections): larger exergy destruction in a downstream unit: case in point is the intercooling in a multi-stage …

Community detection algorithms identify subsets of nodes closely interconnected, which frequently correspond to distinct topics within the document collection.

The first three Communities of related chunks:

Community 1 (51 chunks):

 - Chunk 23: classification is somewhat arbitrary, because some of the rules intrinsically per...
 - Chunk 24: 2.1. General Process-Related Rules 1.Do not use excessively large or excessively...
 - Chunk 25: prescribes that the temperature differences between the heating and heated fluid ...

Community 2 (37 chunks):

 - Chunk 0: A Critical Interpretation and Quantitative Extension of the Sama-Szargut Second ...
 - Chunk 1: “second law guidelines” for design and process engineers, nicknamed at that time...
 - Chunk 2: design of energy conversion systems. The list was rearranged several times, unti...

Community 3 (6 chunks):

 - Chunk 49: panel), its intrinsic technological limitations (in this case, the operating tem...
 - Chunk 42: systems are large, complex and non-linear systems, and a change in the type, qua...
 - Chunk 43: of the other components, be they up or downstream in the conversion chain. While

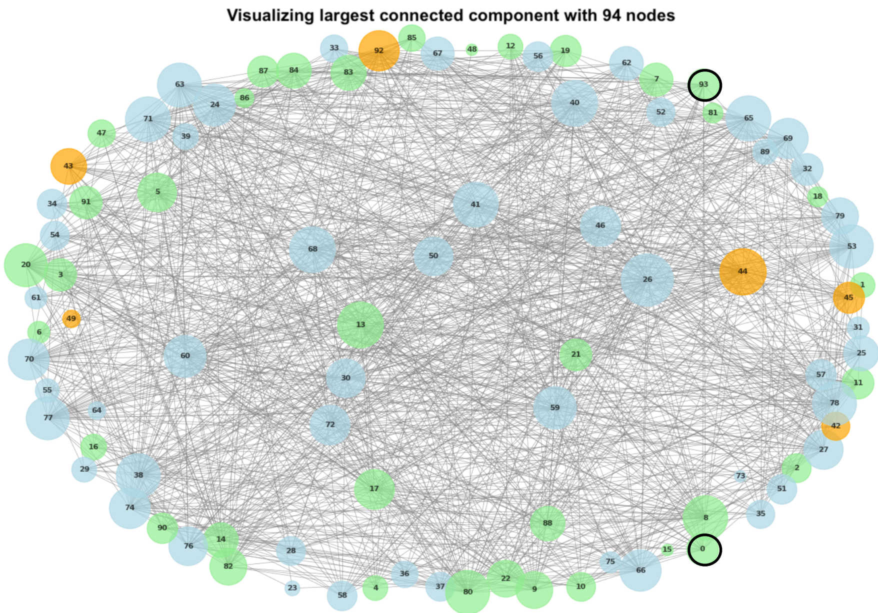

Figure 3: Graph visualizing largest connected components

Figure 3 illustrates the relationship and connections between different sections within the document. Larger nodes indicate a higher number of connections, highlighting their central significance across multiple topics. Different colors represent different communities of related segments.

7 KNOWLEDGE GRAPH-RAG INFORMATION RETRIEVAL

The retrieval process initially identifies relevant data segments (chunks) and subsequently broadens the search through graph-based connections to uncover related information that may exhibit indirect or semantic associations with the query but is linked to relevant chunks. By utilizing the graph, it is possible to identify not only the directly relevant segments when asking a question but also related segments through graph traversal algorithms.

As shown below, Graph RAG efficiently retrieves more (5 directly related chunks and 5 more related from graph connections) contextually relevant documents by effectively understanding the semantic and structural relationships among concepts compared to RAG (only the first three directly connected chunks).

In the example the query is: *What is the impact of the pinch point on heat exchanger network operational efficiency?*

The result has found 5 initial relevant chunks, expanded to 14 chunks using graph connections (chunk descriptions have been truncated by the software and shown as a reference)

Retrieved 10 chunks:

1. [DIRECT] (similarity: 0.802)

Eliminate leaks in pipelines, valves and combustion chambers. Even small leakages of compressed gases, of hot combustion gases or unplanned air i...

> *Compared to RAG:*
> *Result 1 (score: 0.399): Eliminate leaks in pipelines, valves and combustion chambers. Even small leakages of compressed gases, of hot combustion gases or unplanned air i...*

2. [DIRECT] (similarity: 0.800)

Consider the influence of possible changes in the energy management of a part of a system on the exergy destructions in other links of the system. ...

> *Compared to RAG:*
> *Result 2 (score: 0.401): Consider the influence of possible changes in the energy management of a part of a system on the exergy destructions in other links of the system. ...*

3. [DIRECT] (similarity: 0.800)

cannot be eliminated from a heat exchanger, because with vanishing temperature differences the heat transfer area would be infinitely large; the exerg...

> *Compared to RAG:*
> *Result 3 (score: 0.405): cannot be eliminated from a heat exchanger, because with vanishing temperature differences the heat transfer area would be infinitely large; the exerg...*

4. [DIRECT] (similarity: 0.797)

parallel equivalent. Counter-flow heat exchangers and selective membranes display the smallest exergy destructions (see rule 1). Application of parall...

5. [DIRECT] (similarity: 0.796)

influenced not only by thermodynamic factors (like efficiency or specific fuel consumption, source-end use matching etc.) but also-and predominantly-by ...

6. [GRAPH] (similarity: 0.796)

Minimize the throttling of steam or other gases. Throttling introduces a destruction of exergy which could be otherwise recovered, and is usually...

7. [GRAPH] (similarity: 0.795)
Avoid to unduly extend the chain of thermodynamic processes for a certain product. Every additional link in the chain is a real process and intro...

8. [GRAPH] (similarity: 0.787)
the chain of processes and therefore ensures a reduction of the exergy destruction within the entire system (see rule 17). Furthermore, synergic effe...

9. [GRAPH] (similarity: 0.786)
multi-stage compressor in a gas turbine plant: the air stream that enters the combustion chamber is at lower temperature than without intercooling, a...

10. [GRAPH] (similarity: 0.783)
Viscous Loss-Related Rules 13. Exergy destructions due to hydraulic friction or irreversible heat transfer are the larger, the lower is the temp...

CONCLUSIONS

RAG (Retrieval-Augmented Generation) systems integrate information retrieval techniques with natural language generation models to deliver contextually relevant and semantically coherent responses to user inputs. Knowledge graphs confer multiple technical advantages, combining the strenghts of LLMs with external knowledge sources therefore enhancing the efficacy of such systems. The key benefits are outlined as follows:

- Structured knowledge representation and relational contex
- Contextual understanding
- Inferential reasoning
- Knowledge integration
- Explainability and transparency

REFERENCES

1 - Sama, D.A., Qian, S., Gaggioli, R.A., June 1989, *A common-sense 2nd law approach for improving process efficiencies*. In Proceedings of the Thermodynamic Analysis and Improvement of Energy Systems (TAIES '89), Beijing, China, pp. 520–531.

2 - Szargut, J., Sama, D.A. June 1995, *Practical rules of the reduction of energy losses caused by the thermodynamic imperfections of thermal processes*. Proceedings of the 2nd International Thermal Energy Congress, Agadir, Morocco; Volume 2, pp. 782–785.

3 - Sciubba, E., 2014, *Why are some of the Sama-Szargut Second Law rules routinely violated in practical applications?*, 1st International e-Conference on Energies, Energies.

4 - Alexopoulos, P., August 2020, *Semantic Modeling for Data - A practitioner Guide,* O'Reilly, 328 p.

5- Barrosa, J., Webber, J., June 2023, *Building Knowledge Graphs- A practitioner Guide,* O'Reilly, 290 p.

6 - Sciubba, E., 2014, *A Critical Interpretation and Quantitative Extension of the Sama-Szargut Second Law Rules in an Extended Exergy Perspective*, *Energies*, *7(8): 5357-5373*.

7 - Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R., A., Mukherjee, S., Tang, X., He, Q., Hua, Z., Long, B., Zhao, T., Shah, N., Javari, A., Xia, Y., Tang, J., January 2025, *Retrieval-Augmented Generation with Graphs (GraphRAG),* arXiv:2501.00309v2 [cs.IR].

8 - Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., Larson, J., February 2025, *From Local to Global: A GraphRAG Approach to Query-Focused Summarization*, arXiv:2404.16130v2 [cs.CL].
9 - Malkov, Y, A,. Yashunin, D. A., April 2020, *Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs,* IEEE Transactions on pattern analysis and Machine Intelligence, VOL. 42, No. 4, pp 824-836.
10-Yang, Y., Wu, H., Wang, T., Yang, J., Ma, H., Luo, G., March 2025, *Pseudo-Knowledge Graph: Meta-Path Guided Retrieval and In-Graph Text for RAG-Equipped LLM,* arXiv:2503.00309v1 [cs.IR]
11-Hilel, M., Karmim, Y., De Bodinat, J., Sarehane, R., Gillon, A., November 2025, *Building Specialized Software-Assistant ChatBot with Graph-Based Retrieval-Augmented Generation,* arXiv:2511.05297v1 [cs.SE].
12-Shirakawa, M., Arakawa, M., Nakayama, H., October 2010, *Multi-objective optimization system for plant layout design (1st report, genetic algorithm approach for designing layout of plot plan for power plants),* Transactions of the Japan Society of Mechanical Engineers Series C. 76. 2694-2703. 10.1299/kikaic. 76.2694.
13-Shirakawa, M., Arakawa, M., January 2011, *Multi-Objective Optimization System for Plant Layout Design (2nd Report, Application of Particle Swarm Optimization to Automatic Placement and Adjustment of Plant Systems),* TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C. 77. 1609-1620. 10.1299/kikaic. 77.1609.
14-Shirakawa, M,. Arakawa, M., January 2018. *Multi-objective optimization system for plant layout design (3rd report, Interactive multi-objective optimization technique for pipe routing design).* Journal of Advanced Mechanical Design, Systems, and Manufacturing. 12. JAMDSM0053-JAMDSM0053. 10.1299/jamdsm.2018jamdsm0053.
15-Deng, J., L., November 1989, *Introduction to Grey system theory*, The Journal of Grey System, Volume 1, Issue 1, pp 1 - 24
16-Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Osazuwa Ness, R., Larson, J., February 2025, *From Local to Global: A GraphRAG Approach to Query-Focused Summarization,* arXiv:2404.16130.
17- Morwani, D., Zhang, H., Vyas, N., Wu, J., Zou, D., Ghai, U., Foster, D., Kakade, S., April 2025, *How Does Critical Batch Size Scale in Pre-training?, arXiv:2410.21676v4*
18-Maiorano M., Petrucci J., Sciubba E., 2002, *An Expert Assistant for the automatic synthesis of Heat Exchanger Networks*, Proc. of ECOS 2002, July 3-5, Berlin, Germany