Towards an Agentic Smart Research Environment: Hybrid Knowledge Graph Construction via Multi-Level Chunking and Multi-Stage Quality Gating in Research Agentic AI MySRE

1st Muhammad Zaidan Rafat
Dept. Informatics,
Faculty of Science & Technology,
UIN SMH Banten,
Serang, Indonesia
231730016.zaidan@uinbanten.ac.id

2nd Rio Nurtantyana
Research Group HCI-V,
Data and Information Science Center,
National Research and Innovation Agency,
Bandung, Indonesia
rion003@brin.go.id

3rd Ahmad Tabrani
Dept. Informatics,
Faculty of Science & Technology,
UIN SMH Banten,
Serang, Indonesia
Ahmad.tabrani@uinbanten.ac.id

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Scientific and engineering output has grown rapidly worldwide. The National Science Board (2024) reports that "From 2003 to 2022, annual S&E publications increased by roughly a third (36%) in the United States, whereas they increased approximately 10-fold in China and 8-fold in India," while "Gold OA articles increased over 50-fold, from 19,000 articles published in 2003 to 992,000 articles in 2022" [1]. This volume of information creates what Lu et al. (2025) describe as "the exponential growth of scientific literature, with over 7 million articles published annually, [which] exposes a significant bottleneck: the widening gap between unstructured knowledge in texts and its structured representation in KGs" [2]. Consequently, "Traditional approaches to KG enrichment, such as manual curation, are reliable but unsustainable at scale" [2], requiring automated knowledge management systems.

Retrieval-Augmented Generation (RAG) emerged to handle such knowledge-intensive text. As Lewis et al. (2020) state, "Pre-trained neural language models... cannot easily expand or revise their memory, can't straightforwardly provide insight into their predictions, and may produce 'hallucinations'," thus "Hybrid models that combine parametric memory with non-parametric (i.e., retrieval-based) memories can address some of these issues" [3]. However, this basic flat vector approach suffers from context loss. Singh et al. (2026) clarify that "traditional RAG pipelines are typically static and linear, limiting complex multi-step reasoning, deep contextual understanding, and iterative response refinement" [4]. To solve this, the methodology shifts to Agentic RAG, which "integrates autonomous agents directly into the RAG pipeline to enable dynamic retrieval" [4]. For example, in the scientific domain, Lála et al. (2023) built PaperQA because "standard RAG models follow a fixed, linear flow," and they successfully "eliminate these limitations by breaking RAG into modular pieces, allowing an agent LLM to dynamically adjust and iteratively perform steps in response to the specific demands of each question" [5].

If an LLM builds a Knowledge Graph by pure brute-force calculation across documents ($O(N^2)$), it encounters fatal bottlenecks in computation and semantic integrity. Computationally, Huang et al. (2024) prove that "The fundamental issue is that LLMs cannot properly judge the correctness of their reasoning... LLMs struggle to self-correct their responses without external feedback" [6]. Without a Native Vector Index to prune the search space early, $O(N^2)$ scalability is impossible; Sehgal and Salihoğlu (2025) note that modern systems must use a "prefiltering approach" before connecting chunks with structured records like a knowledge graph [7]. Semantically, forcing LLMs to define relations without filters degrades accuracy. Berman et al. (2025) warn that "existing methodologies tend to overlook the generation of edge attributes," and "Without edge features, this information would have to be redundantly stored at the node level, increasing complexity and obscuring transition flow" [8]. To prevent the graph from becoming a tangled 'hairball' of false relations, Hassan et al. (2025) assert that "Evaluation measures the quality of the KG, including its completeness, consistency, accuracy and relevance... By evaluating the KG, errors and inconsistencies can be identified and corrected" [9].

This article proposes the Agentic Smart Research Environment (MySRE) as an architectural solution. MySRE introduces a hybrid extraction model using Multi-Level Chunking and Multi-Stage Quality Gating (K-NN Thresholding, Semantic Coherence, and Domain Compatibility) to bypass $O(N^2)$

Identify applicable funding agency here. If none, delete this.

complexity and filter semantic hallucinations. To demonstrate this superiority, generation quality is measured using Ragas; as Es et al. (2024) explain, Ragas offers "a suite of metrics which can be used to evaluate these different dimensions without having to rely on ground truth human annotations" [10]. For the resulting Knowledge Graph, topological metrics via NetworkX are calculated. Newman (2003) validates this by stating that graph theory "aims to find statistical properties, such as path lengths and degree distributions, that characterize the structure and behavior of networked systems" [11]. Finally, Seo et al. (2022) justify that "As a complement to the shortcomings of current structure-based metrics and data-quality-based assessment techniques, we introduce the structural quality metrics as a measure of knowledge graph quality" [12]. This ensures that MySRE produces structurally precise and highly organized results.

II. II. Related Works

A. Static RAG Limitations and the Illusion of Agent Autonomy

Traditional Retrieval-Augmented Generation (RAG) architectures execute linearly, a structural decision that frequently results in substantial information loss during knowledge extraction. Parametric-only language models naturally struggle to adapt to new facts. Lewis et al. [3] observed that parametric-only models like T5 or BART inherently require continuous retraining to incorporate real-world changes, and without such updates, their predictions degrade into structural hallucinations. While hybridizing parametric memory with non-parametric retrieval mitigates some of these issues, conventional RAG systems remain fundamentally limited. Singh et al. [4] noted that static workflows restrict conventional RAG architectures from executing dynamic, multistep reasoning, a limitation that necessitated the structural shift toward agentic intelligence.

While Agentic RAG provides an alternative to static workflows, search agents currently operate under an illusion of autonomy. Autonomous agents demonstrate competence in summarizing scientific domains. For instance, Lála et al. [5] constructed PaperQA to operate by independently finding relevant papers, gathering texts, and subsequently generating narrative answers. Yet, their outputs remain purely narrative. These agents summarize facts intelligently but operate with structural blindness. They lack the architectural capacity to interconnect scientific facts into a cohesive Knowledge Graph, leaving the generated knowledge flat and isolated.

B. The Probabilistic Paradox: Graph Extraction, Scalability, and Relational Hallucinations

The current shift toward automated graph extraction exposes a probabilistic paradox: science demands deterministic objectivity, whereas language models inherently compute probability distributions. Allowing a language model to build a graph through brute-force computation across raw text leads to catastrophic scalability failures. Lu et al. [2] identified this failure mode, demonstrating that attempting to build knowledge graphs by parsing full text through language models reliably results in relational hallucinations, a complete loss of schema consistency across documents, and an unsustainable quadratic explosion in computational costs.

Delegating logical validation entirely to the language model exacerbates these problems. Huang et al. [6] established that language models cannot self-verify their own logic; they demonstrated that relying on models for self-correction is fundamentally flawed, as performance typically degrades rather than improves when models attempt to fix their own reasoning errors. Without deterministic external constraints, unguided models invent false relationships. Berman et al. [8] confirmed that relational connections are a primary source of structural hallucinations, noting that existing generative frameworks either entirely exclude edge attributes or treat them separately, resulting in models that systematically fail to generate accurate relational features. Consequently, post-filtering methods collapse under immense computational loads. Sehgal and Saliho0̆308glu [7] provided the definitive mathematical case against conventional GraphRAG approaches by proving that computing a specific subset of data prior to executing a nearest-neighbor search is the only reliable method to prevent the system from collapsing. This proves that a prefiltering mechanism is mathematically required before edge extraction.

C. Architectural Resolution: Multi-Level Chunking and Deterministic Gating

To address context loss without incurring quadratic processing costs, our methodology implements a Parent-Child (Multi-Level) chunking mechanism coupled with deterministic thresholding. Retrieving fragmented text segments sacrifices the semantic integrity of the document. Sarthi et al. [13] established the necessity of hierarchical context preservation, proving that retrieving isolated short chunks destroys the holistic understanding of a document, whereas organizing texts into hierarchical trees preserves both contextual and semantic coherence.

However, retrieving larger parent contexts requires strict boundaries to prevent noise injection. Luan et al. [14] analyzed the limitations of dense vectors, demonstrating that fixed-length encodings fail to maintain precise retrieval margins for long documents without the implementation of strict similarity thresholds and sparse-dense hybrid approaches. This finding mandates a strict K-Nearest Neighbors (K-NN) similarity cutoff to filter out noisy data before the graph generation phase.

In scientific domains, vector similarity alone is insufficient. Mysore et al. [15] validated the need for domain-specific coherence filters, finding that in technical literature, true similarity is dictated by matching fine-grained, highly specific aspects of the text rather than relying on generalized vocabulary overlap. This finding mathematically justifies deploying a semantic coherence gate prior to edge formation, ensuring that extracted entities possess specific conceptual alignment rather than superficial word associations.

D. Deterministic Evaluation: Subordinating Linguistic Bias to Mathematical Topology

Evaluating retrieval and generation systems using subjective human baselines introduces inherent linguistic biases. Es et al. [10] legitimized algorithm-driven evaluation by demonstrating that utilizing automated metrics eliminates the subjectivity inherent in human ground-truth annotations, thereby enabling more reliable and objective testing cycles.

When evaluating the resulting Knowledge Graph, conventional approaches disproportionately emphasize scale over structural integrity. Seo et al. [12] criticized conventional methodologies that disproportionately emphasize the size of the graph, arguing instead that evaluating the structural ontology itself is the only mathematically sound approach to determining actual graph quality.

Ultimately, the structural quality of a generated scientific graph must be judged by topological mathematics. Newman [11] formulated the foundation for evaluating massive networks without subjective intervention, establishing that mathematical properties such as path lengths and degree distributions are the definitive arbiters for characterizing the structure and behavior of any complex networked system. These network metrics serve as the ultimate objective arbiter for evaluating the relational quality of our system's Knowledge Graph output.

III. Methodology and System Architecture

This research is anchored in the Design Science Research (DSR) methodology, which provides a rigorous structural foundation for creating and evaluating novel technological solutions. As postulated in foundational literature, "The result of design-science research in IS is, by definition, a purposeful IT artifact created to address an important organizational problem" [16]. Within the context of this study, the MySRE (Multi-Level Chunking & Multi-Stage Quality Gating) architecture represents this definitive IT artifact. The system is engineered explicitly to resolve systemic computational bottlenecks of $O(N^2)$ complexity and the proliferation of false-positive relational edges—commonly referred to as the hairball graph phenomenon—that persistently degrade traditional Retrieval-Augmented Generation (RAG) frameworks.

To ensure the validity and reproducibility of the proposed architecture, the development lifecycle strictly adheres to an established DSR procedural model. Specifically, "The DS process includes six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication" [17]. Adhering to this exact sequence ensures that the engineering phases, spanning from the initial architectural blueprint to the multi-stage validation mechanisms, possess undeniable scientific legitimacy and an internationally recognized problem-solving workflow.

A. MySRE System Architecture Overview

The MySRE architecture is engineered as an autonomous Agentic AI system, designed to operate entirely without human intervention. By deploying a sequential Dual-Pipeline

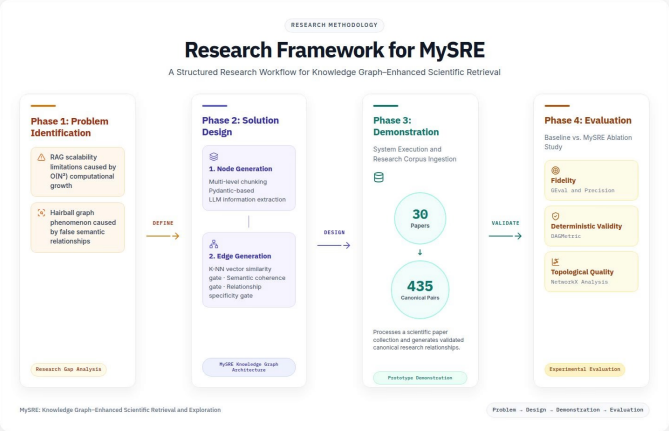

Fig. 1. DSR Methodology Flowchart illustrating the sequential phases applied to the MySRE development lifecycle.

Architecture, the system neutralizes the inherent limitations of conventional Retrieval-Augmented Generation (RAG) frameworks. This segregation of data flow guarantees high-fidelity processing of scientific literature, strictly mapping raw input into an interconnected semantic network.

The first operational phase, designated as the Structural Node Generation pipeline, strictly prohibits the direct ingestion of raw PDF documents into the retrieval engine. Instead, it executes a hierarchical multi-level chunking process to construct a highly precise search index. Following textual decomposition, a Large Language Model extracts five foundational academic pillars: Background, Methodology, Objective, Gap, and Future Work. This extraction phase is governed by a programmatic safeguard termed the Pydantic Armor, which enforces rigid formatting constraints. The culmination of this phase yields structurally pure data nodes, devoid of extraction anomalies or formatting artifacts.

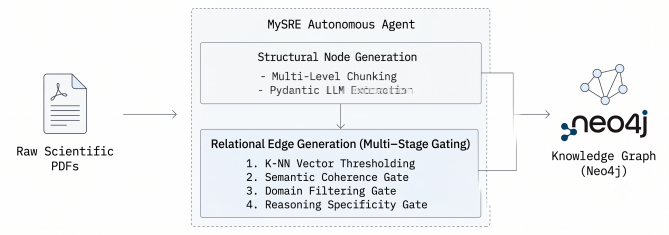

Fig. 2. Blueprint of the MySRE Dual-Pipeline Architecture, illustrating the data flow from raw PDF ingestion to the final Neo4j Knowledge Graph representation.

The second operational phase, the Relational Edge Generation pipeline, is engineered to prevent the formation of arbitrary or irrelevant connections between documents. Before any relational edge is established, candidate connections must survive a rigorous defense mechanism known as Multi-Stage Quality Gating. This defense consists of four sequential filters: K-NN Vector Thresholding for spatial distance selection, a Semantic Coherence Gate to ensure minimal structural alignment across academic pillars, a Domain Filtering Gate for

taxonomic compatibility validation, and a Reasoning Specificity Gate to audit the depth of logical justification. Only relational edges that successfully traverse all four strict gates are permanently materialized and injected into the Neo4j graph database.

B. Structural Node Generation via Multi-Level Chunking

The primary limitation of static Retrieval-Augmented Generation (RAG) frameworks lies in their reliance on naive chunking mechanisms. Splitting text strictly by arbitrary length induces severe context fragmentation, blinding the language model during downstream extraction tasks. Empirical studies have established that passages naturally contain extraneous details capable of distracting both the retriever and the language model during downstream operations. Conversely, while sentence-level indexing offers extreme granularity, sentences often exist as complex and compounded structures that are not self-contained, thereby lacking the necessary contextual information required to accurately judge query-document relevance [18]. To overcome this structural flaw, the MySRE architecture rejects single-layer chunking and institutes a multi-level hierarchical decomposition strategy.

Before textual decomposition occurs, the raw input undergoes an aggressive Data Purification Layer V2. This offline filtration mechanism utilizes deterministic heuristic bounds and rigorous regular expressions to eliminate residual textual noise, such as fractured OCR artifacts, acknowledgments, author affiliations, and hidden bibliographies. The necessity of this pre-processing step is validated by the fact that converting long texts into fixed-length dense vectors carries a major risk of losing essential information. This limitation becomes critically emphasized when source texts are inundated with low-quality, noisy text, invariably resulting in inconsistent retrieval quality [19]. Failing to eradicate these artifacts mathematically distorts cosine similarity proximities in the downstream vector space.

Following purification, the text is decomposed using a Small-to-Big Retrieval architecture, separating the search index from the generation context. The search engine requires short, precise text clusters (Child Chunks) to maintain sharp semantic vectors, whereas the language model demands extensive surrounding text (Parent Chunks) for coherent reasoning. This decoupling strategy is statistically superior, as "even the least performing Sentence Window technique surpasses the best Classic VDB method in retrieval precision" [20]. To protect against silent truncation, the system enforces a strict 1500-character Token Limit Guard. This hard limit aligns with the structural reality of the underlying E5 dense model, which states that "The mE5small, mE5base and mE5large are initialized from the multilingual MiniLM (Wang et al., 2021), xlm-roberta-base (Conneau et al., 2020), and xlm-roberta-large respectively" [21], inherently binding the architecture to an absolute ceiling of 512 tokens.

To navigate the vector space efficiently, MySRE abandons conventional natural language question-and-answer queries in favor of an Anchor-Density Strategy. Dense retrievers are highly reactive to concentrated lexical anchors rather than

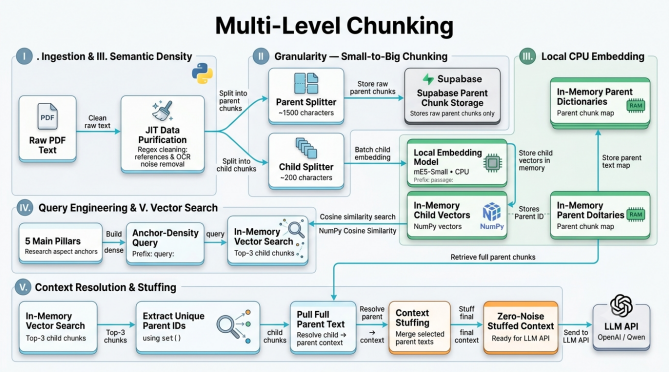

Fig. 3. Diagram illustrating the Multi-Level Chunking architecture, from offline Data Purification V2 to Parent-Child separation and constrained JSON extraction.

diluted conversational grammar. From an information theory perspective, optimizing attention entropy requires increasing information entropy at positions likely to contain key information, while ensuring that the overall total information entropy of the context remains stable [22]. Long narrative queries dilute attention, whereas dense nominal anchors compress semantic signals directly onto critical academic pillars. This transition aligns with the evolution of Agentic RAG systems that "equip LLM agents with retrieval tools, enabling dynamic query formulation" [23]. Furthermore, the retrieval engine bridges lexical alignment prior to conceptual representation, operating on the premise that "semantic similarity of contextualised token embeddings should provide superior retrieval compared to exact string matching. Therefore, MDRAFT is represented by a non-parametric datastore that employs approximate nearest neighbour search to match the (full) current context x to similar contexts in the datastore" [24].

Once the relevant Parent Chunks are retrieved via the anchor-density strategy, the system implements a Context Stuffing mechanism. Instead of extracting the five research pillars individually—which risks logical discontinuity—the unique deduplicated chunks are concatenated into a massive single variable. To preserve document identity, the Raw Document Header (the initial 1500 characters) is injected at the apex of this stuffed context. Feeding this extensive context window into the LLM enables the model to achieve a holistic overview of the document anatomy. Research confirms that "We find that extending the context window of LLMs and providing them with longer retrieved contexts significantly improves downstream performance compared to using short retrieved chunks. The long context enables the model to reason over a broader scope of information, demonstrating the synergistic effect of RAG and long-context LLMs" [25].

Finally, the extraction of the academic pillars is governed by Dual-Layer Constrained Generation. Language models possess an inherent vulnerability to fabrication because traditional supervised fine-tuning methods typically force models to complete every response. Consequently, when presented with queries extending beyond their knowledge boundaries, these

models exhibit a higher likelihood of fabricating content rather than rejecting the prompt, confidently producing falsehoods without an awareness of their own knowledge limits [26]. To neutralize this threat, the system forces a strict JSON schema output, fortified by a Pydantic BeforeValidator. This programmatic armor coerces any hallucinatory conversational phrases—such as apologies or statements regarding missing information—into mathematically safe null strings, ensuring the absolute structural purity of the generated nodes.

C. Relational Edge Generation & Multi-Stage Quality Gating

1) Gate 1 – Statistical K-NN Vector Thresholding: Traditional knowledge graph construction typically employs Cosine Similarity scores with static, hardcoded thresholds, commonly hitting flat marks such as ≥ 0.70. This blind approach ignores the inherent spatial biases of dense vectors, resulting in the Hairball Graph phenomenon where setting the threshold too low produces massive false connections, while setting it too high eradicates essential relations. MySRE introduces an Adaptive Statistical Threshold, mathematically calibrating the noise floor dynamically according to the actual data population distribution through the Three Sigma Rule of Thumb.

The system evaluates semantic distance by calculating the Cosine Similarity between document vectors:

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

For any incoming query document q, the engine initially searches the database D to fetch a candidate set S bounded by a maximum limit $K = 5$. This hyperparameter (`VECTOR_TOP_K`) strictly restricts the maximum candidate volume retrieved per query to prevent system memory overload (OOM):

$$S = \text{KNN}(q, D, K) \quad (2)$$

To dynamically identify anomalies, the system periodically samples a defined population of $N = 500$ random pairs (`SAMPLE_SIZE`) from the database during the distribution curve calibration process to compute the population mean (μ) and sample standard deviation (σ). An execution barrier is then constructed using the Z-Score multiplied by an academic pillar-specific sensitivity weight (M_{pilar}):

$$Z_{score} = \mu + 3\sigma \quad (3)$$

$$\text{Final Threshold}_{pilar} = Z_{score} \times M_{pilar} \quad (4)$$

Candidates within the set S must survive this brutal cutoff. The mathematical validity of this approach is founded upon the 3σ rule, which states that for any unimodal distribution, the probability of a random variable falling outside the interval $[\mu - 3\sigma, \mu + 3\sigma]$ is at most 5% and specifically 0.27% for an exact normal distribution [27]. Consequently, the system operates on the statistical anomaly detection principle, which assumes that normal data instances occur in high probability regions of a stochastic model, leading to the labeling of an observation as an anomaly if it possesses a low probability of being generated by that model [28].

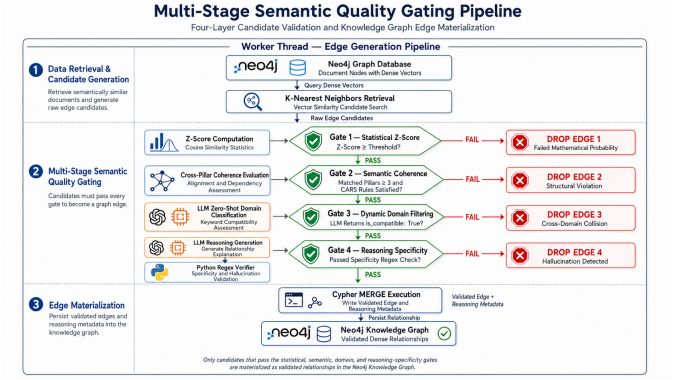

Fig. 4. Architecture of the Multi-Stage Quality Gating system, visualizing the sequence of the Statistical K-NN Vector Thresholding, Semantic Coherence Gate, Dynamic Domain Filtering Gate, and Reasoning Specificity Gate.

The implementation of asymmetric hyperparameter weights (Multipliers) addresses the architectural vulnerability of dense vectors. Research confirms that fixed-length dense encodings can struggle with the precise retrieval of long documents due to complex connections between the encoding dimension, document length, and the structural margin separating relevant from irrelevant documents [14]. Therefore, while the Background and Methodology pillars receive a 1.0x multiplier, the Objective and Future Work pillars are tightened by 5% (1.05x). The Gap pillar undergoes the most brutal filtration at 1.15x, demanding near-identical vector proximity to establish a valid relation.

TABLE I
ASYMMETRIC HYPERPARAMETER WEIGHTS (MULTIPLIERS) ACROSS ACADEMIC PILLARS

Parameter	Value	Function & Logical Rationale
`Multiplier` (Background)	1.0x	No penalty. Background sections are highly generalized and formulaic.
`Multiplier` (Methodology)	1.0x	No penalty. Procedural and tooling methodologies are heavily repetitive.
`Multiplier` (Objective)	1.05x	Tightened by 5%. Research objectives must be sharp to be considered intersecting.
`Multiplier` (Future Work)	1.05x	Tightened by 5%. Future projections require high lexical synchronization.
`Multiplier` (Gap)	**1.15x**	**Most Brutal Filtration**. The gap determines research originality. The semantic vector must be nearly identical to form a relation (tightened by 15%).

This exact architecture mirrors state-of-the-art implementations where researchers construct Reasoning Knowledge Graphs and eliminate flawed steps that deviate from the consensus by utilizing statistical filtering techniques like z-score filtering [29].

2) Gate 2 – Semantic Coherence Gate: Surviving the distance calculation in Gate 1 does not guarantee structural alignment. Traditional vector search architectures generate false positives by allowing relations based on single-pillar matches. For instance, two papers utilizing the Random Forest

algorithm will register identical Methodology vectors, even if one predicts weather and the other forecasts heart failure. To counter this, MySRE implements the Semantic Coherence Gate, a verification layer enforcing cross-pillar validation to ensure relationships adhere strictly to academic rhetorical rules.

To maintain optimal execution speed without relying on the heavy $O(N^2)$ quadratic distance calculations required by optimal transport frameworks, MySRE reduces complexity through an absolute $O(1)$ Majority Voting heuristic. The system parameter requires a minimum consensus of three matched pillars (`COHERENCE_MIN_MATCHED_PILLARS = 3`). Since the total ontology consists of five pillars, demanding three matches requires the consensus to exceed 50%. This cross-dimensional requirement reflects the empirical reality that scientific papers describe multifaceted arguments and ideas, suggesting that models capable of matching specific aspects can better capture overall document relatedness by flexibly aggregating over fine-grained sentence-level aspect matches [15].

Furthermore, these multi-aspect validations are bound by the CARS (Creating a Research Space) framework [30]. Under this rhetorical doctrine, the description of methodology acts solely as an operational instrument subordinate to the fulfillment of the research objective designed to resolve a specific gap. Translating this postulate into executed code, the system enforces a strict Boolean dependency matrix. A Methodology match is instantly nullified if it is not accompanied by an intersection in the Objective or Gap pillars, obliterating the formation of false algorithm clusters. Similarly, a shared Objective or Gap is deemed void without a corresponding procedural approach or defined research space.

TABLE II
BOOLEAN DEPENDENCY MATRIX BASED ON THE CARS FRAMEWORK

Anchor Pillar	**Absolute Dependency Requirement (At least one)**	**Ontological Enforcement Description**
Methodology	Must be accompanied by **Objective** or **Gap**	Nullifies isolated methodology matches to eliminate the formation of false algorithm clusters.
Objective	Must be accompanied by **Background**, **Methodology**, or **Gap**	Research objectives are merely verbs without supporting execution instruments (Methodology) or a localized problem (Gap/Background).
Gap	Must be accompanied by **Objective** or **Methodology**	Sharing a problem constraint is invalid unless both papers deploy a similarly structured execution or objective approach.
Future Work	Must be accompanied by **Objective**, **Methodology**, or **Gap**	Future trajectory suggestions stem from current limitations; similarities are only valid if executed within allied problem perimeters.

3) Gate 3 – Dynamic Domain Filtering Gate: Dense vectors evaluate text strictly via mathematical probability and spatial proximity, leaving them inherently blind to scientific taxonomy. Such cross-domain semantic collisions result in artificial cosine similarity spikes across completely divergent scientific disciplines. Instead of executing highly expensive computational fine-tuning on the base embedding model, MySRE shifts this validation burden to the generative Large Language Model, deploying it as a Zero-Shot Taxonomy Judge.

The generative model is restricted to analyzing only the specific research keywords extracted from the source and target nodes, rather than processing the entire document text. Driven by the explicit system prompt acting as an expert academic domain classifier, the model evaluates latent semantic meaning, normalizes multi-language taxonomy concepts, and resolves technical abbreviations. To protect backend automation from parsing errors, the model is strictly coerced to output a JSON object returning a pure Boolean `is_compatible` variable and a sub-ten-word `reason` string. A negative Boolean trigger results in the immediate truncation of the relational edge. This architectural choice is validated by literature confirming that existing medical LLMs are typically either pre-trained or directly aligned to specialized domains by utilizing Zero-shot Prompting, where only task descriptions are allowed to be entered without past data examples [31].

4) Gate 4 – Reasoning Specificity Gate: Large Language Models exhibit a severe vulnerability to sycophancy and epistemic blindness. As completion machines, these models possess a natural tendency to fabricate affirmative narratives rather than reject instructions when analyzing uncorrelated data [26]. Relying blindly on an LLM to formulate relational justifications would pollute the knowledge graph with generic, hallucinatory connections stating that the texts simply discuss similar topics.

MySRE counteracts this through the Reasoning Specificity Gate. The model is forced to execute a generation-time correction by producing a textual justification for the relationship. However, the system explicitly rejects the notion of LLM self-correction. Research unequivocally proves that LLMs struggle to self-correct their responses without external feedback, and their performance frequently degrades after self-correction attempts. Instead, the implementation utilizes a code executor to serve as the perfect verifier to judge correctness [6]. This external verification relies on deterministic Regular Expressions and keyword matching. Any reasoning text containing generic platitudes is subjected to a hard drop. Furthermore, the system implements an Atomic Facts doctrine, demanding the presence of specific operational passwords—such as dataset, metrics, or accuracy. Computing factual precision involves breaking generations into atomic facts containing isolated pieces of information, acknowledging the inherent trade-off between absolute precision and total recall [32]. Failure to manifest these concrete scientific instruments triggers the instant eradication of the candidate edge.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] National Science Board, “The State of U.S. Science and Engineering 2024,” *Science and Engineering Indicators 2024*, NSB-2024-3, 2024.
[2] Y. Lu, W. Wu, X. Zhao, R. Peng, and J. Wang, “KARMA: Leveraging Multi-Agent LLMs for Automated Knowledge Graph Enrichment,” *arXiv preprint arXiv:2502.06472*, 2025.
[3] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
[4] A. Singh *et al.*, “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG,” *arXiv preprint arXiv:2501.09136*, 2026.
[5] J. Lála *et al.*, “PaperQA: Retrieval-Augmented Generative Agent for Scientific Research,” *arXiv preprint arXiv:2312.07559*, 2023.
[6] J. Huang *et al.*, “Large Language Models Cannot Self-Correct Reasoning Yet,” *International Conference on Learning Representations (ICLR)*, 2024.
[7] G. Sehgal and S. Salihoğlu, “NaviX: A Native Vector Index Design for Graph DBMSs With Robust Predicate-Agnostic Search Performance,” *CIDR*, 2025.
[8] N. Berman, E. Kosman, D. D. Castro, and O. Azencot, “Reviving Life on the Edge: Joint Score-Based Graph Generation of Rich Edge Attributes,” *Transactions on Machine Learning Research*, 2025.
[9] H. Hassan, S. Elayidom, M. R. Irshad, and C. Chesneau, “A detailed analysis into evaluation metrics for knowledge graph evaluation,” *International Journal of Data Science and Analytics*, vol. 20, pp. 6933–6972, 2025.
[10] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 150–158, 2024.
[11] M. E. J. Newman, “The structure and function of complex networks,” *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
[12] S. Seo, H. Cheon, H. Kim, and D. Hyun, “Structural Quality Metrics to Evaluate Knowledge Graph Quality,” *arXiv preprint arXiv:2211.10011*, 2022.
[13] P. Sarthi *et al.*, “RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval,” *International Conference on Learning Representations (ICLR)*, 2024.
[14] Y. Luan *et al.*, “Sparse, Dense, and Attentional Representations for Text Retrieval,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329–345, 2021.
[15] S. Mysore *et al.*, “Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity,” *Conference of the North American Chapter of the Association for Computational Linguistics*, 2022.
[16] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
[17] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.
[18] T. Chen *et al.*, “Dense X Retrieval: What Retrieval Granularity Should We Use?,” *EMNLP*, 2024.
[19] H. Tan *et al.*, “QAEA-DR: A Unified Text Augmentation Framework for Dense Retrieval,” *IEEE TKDE*, 2025.
[20] M. Eibich *et al.*, “ARAGOG: Advanced RAG Output Grading,” *arXiv preprint arXiv:2404.01037*, 2024.
[21] L. Wang *et al.*, “Multilingual E5 Text Embeddings: A Technical Report,” *arXiv preprint arXiv:2402.05672*, 2024.
[22] Y. Wang *et al.*, “BEE-RAG: Balanced Entropy Engineering for Retrieval-Augmented Generation,” *arXiv preprint arXiv:2508.05100*, 2025.
[23] E. Lumer, “Rethinking Retrieval: From Traditional Retrieval Augmented Generation to Agentic and Non-Vector Reasoning Systems in the Financial Domain,” *arXiv preprint arXiv:2511.18177*, 2025.
[24] M. Gritta *et al.*, “DReSD: Dense Retrieval for Speculative Decoding,” *Findings of the ACL*, 2025.
[25] P. Xu *et al.*, “Retrieval Meets Long Context Large Language Models,” *ICLR*, 2024.
[26] L. Huang *et al.*, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Computing Surveys*, 2024.
[27] F. Pukelsheim, “The Three Sigma Rule,” *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.
[28] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
[29] Y. Weng *et al.*, “Correct Prediction, Wrong Steps? Consensus Reasoning Knowledge Graph for Robust Chain-of-Thought Synthesis,” *arXiv preprint arXiv:2604.XXXXX*, 2026.
[30] J. Swales, “Genre Analysis: English in Academic and Research Settings,” *Cambridge University Press*, 1990.
[31] H. Zhou *et al.*, “A Survey of Large Language Models in Medicine: Progress, Application, and Challenge,” *arXiv preprint arXiv:2311.05112*, 2023.
[32] S. Min *et al.*, “FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation,” *EMNLP*, 2023.