mathematics

MDPI

Article

Measuring Semantic Coherence of RAG-Generated Abstracts Through Complex Network Metrics

Bady Gana[1,*], Wenceslao Palma [1], Freddy A. Lucay [1,2,*], Cristóbal Missana [1], Carlos Abarza [1] and Hector Allende-Cid [1,3,4]

[1] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Avenida Brasil 2241, Valparaiso 2362807, Chile; wenceslao.palma@pucv.cl (W.P.); cristobal.missana@pucv.cl (C.M.); carlos.abarza.s@mail.pucv.cl (C.A.); hector.allende-cid@iais.fraunhofer.de (H.A.-C.)
[2] Escuela de Ingeniería Química, Pontificia Universidad Católica de Valparaíso, Avenida Brasil 2162, Valparaiso 2362807, Chile
[3] Knowledge Discovery, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[4] Lamarr Institute for Machine Learning and Artificial Intelligence, 53115 Dortmund, Germany
* Correspondence: bady.gana@pucv.cl (B.G.); freddy.lucay@pucv.cl (F.A.L.)

Abstract

The exponential growth of scientific literature demands scalable methods to evaluate large-language-model outputs beyond surface-level fluency. We present a two-phase framework that separates generation from evaluation: a retrieval-augmented generation system first produces candidate abstracts, which are then embedded into semantic co-occurrence graphs and assessed using seven robustness metrics from complex network theory. Two experiments were conducted. The first varied model, embedding and prompt configurations, achieved results showing clear differences in performance; the best family combined `gemma-2b-it`, a prompt inspired by chain-of-Thought reasoning, and `all-mpnet-base-v2`, achieving the highest graph-based robustness. The second experiment refined the temperature setting for this family, identifying $\tau = 0.2$ as optimal, which stabilized results (sd $= 0.12$) and improved robustness relative to retrieval baselines ($\Delta E_G = +0.08$, $\Delta\rho = +0.55$). While human evaluation was limited to a small set of abstracts, the results revealed a partial convergence between graph-based robustness and expert judgments of coherence and importance. Our approach contrasts with methods like GraphRAG and establishes a reproducible, model-agnostic pathway for the scalable quality control of LLM-generated scientific content.

Keywords: RAG; complex networks; semantic graphs; weighted kappa; graph robustness

MSC: 68R10

check for updates

Academic Editors: Adolfo Ballester-Bolinches and Michele Bellingeri

Received: 5 September 2025
Revised: 16 October 2025
Accepted: 20 October 2025
Published: 31 October 2025

Citation: Gana, B.; Palma, W.; Lucay, F.A.; Missana, C.; Abarza, C.; Allende-Cid, H. Measuring Semantic Coherence of RAG-Generated Abstracts Through Complex Network Metrics. *Mathematics* **2025**, *13*, 3472. https://doi.org/10.3390/math13213472

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large language models (LLMs) increasingly generate scientific text, including structured outputs such as abstracts. However, evaluating the semantic quality of these machine-generated abstracts remains a significant challenge. Standard metrics like ROUGE, BERTScore, or factual consistency checks often focus on local similarity or correctness, but they do not assess whether a generated abstract meaningfully integrates into the broader discourse of a scientific field.

Scientific writing exhibits distinct co-occurrence patterns of terms, which reflect the conceptual organization of knowledge within a domain [1]. For example, specific technical

terms tend to appear together in abstracts belonging to the same subfield, and their co-occurrence structures form a semantic network that implicitly encodes the field's knowledge space. When a new abstract aligns well with the field, its terms should naturally fit into this existing network; conversely, poorly integrated or semantically inconsistent abstracts exhibit anomalous patterns.

We leverage this insight by proposing a network-based evaluation method: we assess how well a generated abstract integrates into a semantic co-occurrence network constructed from a representative corpus of scientific documents. A well-constructed abstract typically exhibits a conceptual network of terms and relationships that is coherent (a well-connected core of ideas), resilient to the loss terms (not reliant on a single "hinge" word), efficient (with key ideas reached in a few steps), and balanced (without fragmentations or thematic "islands"). By embedding the abstract into the semantic network and analyzing global and local graph properties via robustness metrics such as spectral radius and efficiency, we obtain holistic signals of its semantic coherence, topical relevance, and alignment with the field.

Our approach builds upon prior work in which co-occurrence networks were used to distinguish human-written from machine-generated text [1]. We extend this idea to scientific abstracts and, critically, we propose a two-phase pipeline that deliberately separates generation from evaluation. In our framework, the LLM produces an abstract without any graph constraints, after which the text is evaluated by measuring how well it integrates into a semantic network of authentic documents. This design contrasts directly with methods such as GraphRAG [2,3], which inject graph structure during the generation process itself. By decoupling the two phases, our contribution is not to improve generation quality per se but to provide an interpretable and reproducible evaluation layer that captures semantic coherence beyond standard similarity metrics.

This methodology is inspired by ideas from scientometrics and educational technology. In scientometrics, bibliometric networks are used to assess the novelty or impact of research outputs by analyzing their position within citation networks [4]. Similarly, in educational settings, student-generated concept maps viewed as graphs are evaluated via network metrics to gauge understanding [5]. By analogy, we treat the LLM-generated abstract as a node in the semantic network whose structural role reveals its integration and potential contribution to the knowledge space.

Our study contrasts with alternatives that directly extract structured knowledge (e.g., triples or concept graphs) from generated text. While such methods focus on factual consistency, they may overlook the emergent semantic coherence of the text as a whole. In contrast, our network-based evaluation explicitly targets semantic integration: we ask whether the generated abstract fits naturally within the body of retrieved knowledge.

In this paper, we present our two-phase evaluation framework and demonstrate its application to LLM-generated scientific abstracts. We show that network-based metrics provide complementary insights beyond traditional text similarity measures, offering a promising direction for the holistic evaluation of machine-generated scientific content.

Despite these advances, important gaps remain in the literature. Existing similarity-based metrics such as ROUGE, BERTScore, or factuality checks assess local overlap but do not capture global semantic integration [6–8]. GraphRAG approaches [3], in turn, embed graph structure into the generation process, yet there is no reproducible framework that leverages graphs as external, post-generation evaluators [9]. Moreover, little evidence exists as to whether graph-theoretic robustness metrics align with human expert judgments of coherence and importance [8,9]. These gaps motivate the following research questions and hypotheses.

To make our objectives explicit, we formulate the following research questions:

RQ1: *Can graph-theoretic robustness metrics capture the semantic coherence of abstracts generated via retrieval-augmented generation (RAG)?*

RQ2: *Is there an observable alignment, even if partial, between these graph-based metrics and human expert judgments of coherence and importance?*

These questions guide our empirical analysis and provide a concrete basis for testing the contribution of graph-theoretic evaluation in comparison to existing textual metrics.

Based on these research questions, we propose the following hypotheses:

H1: *Abstracts with higher graph-theoretic robustness (e.g., global efficiency, spectral radius, algebraic connectivity) will be perceived by experts as more semantically coherent.*

H2: *The configuration that maximizes robustness metrics will also correspond to higher levels of human agreement on coherence and importance.*

These hypotheses connect our graph-based evaluation framework to human judgments, establishing testable claims that extend beyond standard similarity metrics.

Main contributions.

- We propose a two-phase framework that separates generation from evaluation, in which a simple retrieval-augmented generation (RAG) system produces scientific abstracts, and their semantic coherence is assessed independently using graph-theoretic analysis.
- Each abstract is modeled as a semantic co-occurrence network, characterized through seven robustness metrics (e.g., global efficiency, spectral radius, algebraic connectivity), providing interpretable fingerprints of thematic coherence.
- We conduct a comprehensive experimental study across multiple LLMs, embeddings, and prompting strategies, showing that optimal configurations maximize graph robustness while also yielding higher human inter-rater agreement (weighted κ).

The rest of the paper is organized as follows. Section 2 reviews the background and related work. Section 3 introduces the proposed framework, methodology, and experimental setup. Section 4 presents the results and analysis, highlighting the main findings. Section 5 concludes the study and outlines limitations and directions for future research. Appendix A provides additional information.

2. State of the Art

This section reviews the current landscape in RAG and complex graph-based methods for text analysis and evaluation.

2.1. Retrieval-Augmented Generation

RAG enhances LLMs by integrating external retrieval mechanisms, enabling more accurate and context-aware text generation [10]. A RAG system typically uses a retriever to fetch relevant documents from an external source, followed by a generator (e.g., a seq2seq model) that conditions its output on both the original query and the retrieved content [11]. This approach mitigates issues like hallucinations and fixed knowledge limitations in standalone LLMs [12].

RAG has been successfully applied to tasks such as open-domain question answering [13], knowledge-grounded dialogue [14], fact-based summarization [2], and specialized domains

like clinical report generation [11,15]. By incorporating structured data from knowledge graphs (KGs), RAG systems can further improve factual grounding and relevance [16].

In our context, RAG is used to generate scientific abstracts from retrieved literature, and only after generation do we evaluate the text within a network of related documents. This post-generation stance contrasts with GraphRAG [17], for which a corpus-level graph is constructed and then used during retrieval and summarization to steer what is generated. In GraphRAG, the graph is part of the generation mechanism, whereas here, the graph serves as an external evaluator of semantic fit and coherence. The two lines of work are, therefore, complementary: GraphRAG aims to improve generation by structuring context, while our approach focuses on lightweight, model-agnostic quality control after the fact.

2.2. Complex Graph Networks

Complex networks, characterized by properties like a small-world structure [18] and heterogeneous degree distributions, have long been used to represent semantic relationships in language data [19]. In text analysis, nodes may represent words, concepts, or documents, while edges capture semantic relations such as co-occurrence or similarity.

Semantic and document-level networks can be constructed by linking nodes based on shared keywords or embedding-based similarity (e.g., cosine similarity using TF–IDF or SciBERT vectors) [20]. These graphs reveal topical clusters and central concepts, providing insights into the semantic organization of a corpus.

Recent frameworks classify LLM–KG integration into the following: (i) KG-enhanced LLMs (injecting KG knowledge during training/inference), (ii) LLM-augmented KGs (using LLMs to expand or refine KGs), and (iii) synergized models combining both [21].

Our approach diverges by using RAG to generate content first and only then incorporating it into a complex semantic graph for analysis [22].

2.3. Semantic Graphs for Post-Generation Evaluation

Graph-based representations provide a unified lens to assess whether a machine-generated abstract is coherent in itself and well aligned with related literature. We consider two complementary levels reported in prior work. At the document level, abstracts are nodes and edges reflect conceptual overlap, yielding a corpus network where well-integrated texts form dense neighborhoods and outliers remain weakly connected [23–25]. At the word level, each abstract induces a co-occurrence network whose topology has long been linked to linguistic structure (small-world, scale-free patterns) and textual quality [26–28]. Early evidence showed that graph metrics can distinguish generated from human text [1], and subsequent variants enriched co-occurrence graphs with embedding-based edges and pruning backbones to improve discriminability and interpretability [29–32].

Methodological refinements across this literature converge on a few robust defaults. For co-occurrence graphs, short sliding windows (2–5 tokens) capture syntagmatic relations, while sentence/document windows capture broader topical links; dependency contexts offer a syntactic alternative [33,34]. To mitigate frequency bias, association measures such as PMI/PPMI (or NPMI) are commonly preferred over raw counts [35,36]. On the corpus side, similarity thresholds or statistical backbones retain salient ties and reduce noise [30]. Together, these choices support graph metrics, e.g., efficiency, connectivity, conductance, spectral radius/gap, as informative summary statistics of semantic organization.

In line with a post-generation stance, we use word-level co-occurrence graphs to obtain a within-abstract structural fingerprint and compare it against the reference profile induced via the N retrieved documents (document level providing the contextual baseline). This synthesis leverages the strengths of both views: document networks indicate align-

ment/outlierness in the literature, while co-occurrence topology captures internal semantic organization. Our results show that configurations yielding robust word-level topology also align better with expert judgments, supporting graph metrics as model-agnostic proxies for thematic coherence in RAG-generated abstracts.

3. Methodology

This section describes the methodological framework adopted in this study, detailing the data sources, processing pipeline, and evaluation procedures. The objective is to provide a transparent and reproducible account of how RAG was combined with complex network analysis to assess the structural and semantic quality of generated abstracts. We first describe the data collection process and the construction of the experimental corpus. Next, we outline the end-to-end pipeline, including embedding generation, retrieval, and LLM summarization. The subsequent subsections explain how co-occurrence graphs were built, which network metrics were selected, and how they were used to quantify robustness and coherence. Finally, we present the design of the experimental setup and the expert-based evaluation protocol, establishing a systematic foundation for comparing model configurations and interpreting results. All experiments were executed on a dedicated server equipped with 128 GiB of RAM, an NVIDIA GeForce RTX 4080 (16 GiB VRAM), and a 1 TB SSD; complete hardware and software specifications are provided in Table A2.

3.1. Data Collection

A scientific document corpus was constructed using metadata exported from the Scopus database [37]. The query used for data retrieval was as follows and is presented in Listing 1.

Listing 1. The query targets documents containing the phrase "mineral processing" in the title, abstract, or keywords (TITLE-ABS-KEY), restricted to open-access publications (OA, "all").

```
TITLE-ABS-KEY (mineral processing) AND (LIMIT-TO (OA, ‘‘all’’))
```

The following metadata fields from the exported CSV were used:

1. Title:document title.
2. DOI (digital object identifier): unique and persistent document identifier.
3. Abstract: summary of the document's content.

To refine the dataset, a filter was applied to include only open access publications. This ensures that all documents are readily accessible, promoting transparency and reproducibility.

3.2. Pipeline

The pipeline (see Figure 1) begins with the ingestion of metadata exported from Scopus. Each abstract and the user's query are converted into high-dimensional vectors using a pretrained embedding model (see Section 3.2.1); these vectors are indexed in a vector database to enable efficient semantic retrieval. The query embedding retrieves the N most similar abstracts, and both the abstracts and the original query are subsequently inserted into a prompt template. This prompt is then sent to an instruction-tuned LLM, which returns a Proposed Document (synthetic summary) and the list of the N retrieved abstracts.

Before network modeling, each text (the N abstracts and the proposed summary) is normalized, stripped of punctuation and stopwords, and tokenized while preserving original word order, ensuring consistency in graph construction. A word co-occurrence graph is then generated for each text using a sliding window of size three (see Section 3.3); the resulting set comprises N graphs from the source abstracts plus one for the proposed summary.

Seven metrics are computed on each graph to capture robustness, efficiency, and connectivity: natural connectivity, conductance [38], spectral radius, spectral gap, average edge betweenness, global efficiency, and algebraic connectivity. This intentionally simple RAG design minimizes tunable components and preserves end-to-end traceability, facilitating the reproducibility and interpretability of results.

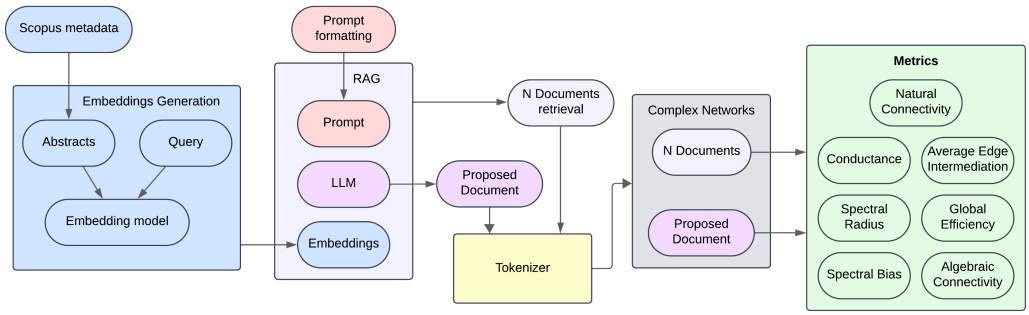


Figure 1. Full pipeline of the RAG system used in this study. The process begins with metadata extraction from Scopus and embedding generation, followed by the retrieval of the top N most similar abstracts, the synthesis of a proposed abstract using an LLM, and the subsequent construction of co-occurrence networks. Seven structural metrics are computed on each graph to compare the semantics of the generated summary with those of the source abstracts.

As part of our two-phase evaluation framework, we employ a simple RAG system (simple/naive RAG) [39] as the text generation component. The choice of a simple RAG system is intentional: the goal of this work is not to optimize the generation process itself but to assess the semantic quality of the resulting abstracts through complex network metrics. While more advanced RAG architectures exist, incorporating multi-stage pipelines, sophisticated reasoning, and enhanced retrieval techniques, this study does not aim to improve RAG performance per se but, rather, to use it as a means for generating abstracts to be evaluated via network-based analysis.

3.2.1. Embedding Generation

The first stage, referred to as embedding generation (Figure A1), involves transforming the textual content of the abstracts into high-dimensional numerical vector representations. Given hardware constraints that prioritized heavier LLMs for later stages, we opted for comparatively smaller and more efficient embedding models.

Two main models from Hugging Face were selected:

- `sentence-transformers/all-mpnet-base-v2`: A small and efficient model designed primarily for information retrieval and short text clustering (384 tokens) in English. It produces 768-dimensional embeddings. This model was chosen for its strong balance between performance and size, making it suitable for resource-constrained environments [40].
- `dunzhang/stella_en_400M_v5`: A relatively small model with 400 million parameters, based on `Alibaba-NLP/gte-Qwen2-1.5B-instruct` and trained using Matryoshka Representation Learning (MRL) [41]. It supports English texts up to 512 tokens and generates 1024-dimensional embeddings. This model offers high-quality representations without the computational burden of larger models [42].

The resulting vector representations are indexed and stored in a vector database to enable fast and efficient retrieval in the subsequent stages of the system [43,44].

3.2.2. RAG

The second stage corresponds to the RAG system (Figure A1). The process begins when a user submits a query (see Listing 2). This query is transformed into an embedding using the same model employed in the previous stage. The resulting query embedding is used to search the vector database and retrieve the top N abstracts most semantically similar to the query.

Listing 2. Example query submitted to an LLM within the RAG system. This prompt poses a question to the LLM, which generates an answer in the form of an abstract.

```
(‘‘What are tailings, and how do environmental,
chemical, and geotechnical factors influence
sustainable tailings management in mineral
processing operations?’’)
```

Based on these similarity scores, the top N abstracts most similar to the query are selected and passed to the LLM, which generates a proposed abstract. The resulting set of abstracts (the N retrieved and the generated one) is then passed to the next stage of the pipeline for graph construction and metric computation.

3.2.3. Large Language Models

Three instruction-tuned language models (instruct models) were selected. These models are optimized to understand and execute specific commands or queries, making them well suited for tasks such as answer generation in RAG systems:

1. `meta-llama/Llama-3.2-3B-Instruct`: Part of Meta's Llama family, this 3-billion-parameter (3B) model is designed to follow instructions. Its relatively small size makes it efficient for deployment in resource-constrained environments without significantly compromising response quality in focused tasks [45].
2. `Qwen/Qwen2.5-3B-Instruct`: Developed by Alibaba Cloud, this 3B model is also instruction-tuned. Qwen is known for its strong performance across a variety of language tasks, often outperforming larger models. Its inclusion enables a direct comparison with Llama-3.2-3B-Instruct, given their similar size and purpose, which is useful for evaluating the performance of different LLM architectures within the same parameter class [46,47].
3. `google/gemma-2b-it`: This Google model, with 2 billion parameters, is the most size-comparable *instruct* version within the Gemma family. Built upon the same research as the Gemini models, Gemma is designed to be lightweight and efficient, making it suitable for local or resource-limited deployments. Despite its smaller size, it performs competitively on reasoning and well-scoped generation tasks [48].

To evaluate each model's summarization capabilities, three prompts were designed with increasing levels of complexity and specificity (full prompt texts are available in Table A1). Each prompt represents a different prompting strategy to examine how instruction formulation affects output quality [49–51].

- Prompt A: This prompt employs the most direct and basic strategy, known as zero-shot prompting. It provides only the retrieved abstracts and a clear instruction to generate a new abstract. The aim is to establish a performance baseline by assessing the model's inherent ability to synthesize information without additional guidance. It serves as the control condition against which more advanced techniques are compared.

- Prompt B: This version introduces the instruction "take your time before answering." It is inspired by Chain-of-Thought (CoT) prompting, which aims to improve reasoning by encouraging the model to reflect before generating an answer [52]. The prompt implicitly guides the model to (1) identify key points, (2) organize them logically, and (3) synthesize a coherent summary. Known as Zero-shot-CoT, this approach enhances fidelity and structure without requiring exemplars, simply by reformulating the instruction to promote deliberate processing [53].
- Prompt C: This prompt uses a role-assignment strategy, instructing the model to act as a "postdoctoral researcher." The goal is to align the model's output with an expert-level tone, terminology, and analytical perspective [54]. By adopting this role, the model is expected not only to summarize content but to do so with the rigor, structure, and stylistic conventions of academic discourse. This strategy provides deep contextual framing to elicit more sophisticated and domain-appropriate outputs [55].

Finally, the texts of the N retrieved abstracts, along with the original user query, are formatted using one of the three defined prompts. This structured prompt is submitted to the LLM, which returns two outputs: (1) a Proposed Abstract, synthesizing the relevant information, and (2) the set of N retrieved abstracts.

3.3. Complex Network Construction

3.3.1. Text Preprocessing

After retrieving the N abstracts along with the one generated via the LLM, each abstract underwent a preprocessing stage. All text was lowercased, punctuation was removed, and the content was split into words using spaces as delimiters. Each term was thus treated as an independent token while preserving the original word order. Subsequently, low-informative terms were removed using the stopword list provided via `nltk` [56]. An inherent limitation of this preprocessing approach is that it treats each lexical item as a single node, disregarding potential polysemy or domain-specific sense variations. In highly specialized contexts, the same term may carry distinct meanings (e.g., "flotation" in chemistry versus process engineering), potentially distorting co-occurrence patterns. While this effect is partly mitigated by using domain-restricted corpora, explicit sense disambiguation remains an open challenge addressed in the Limitations section. Several recent studies propose methods to mitigate this issue through word sense induction [57], multi-sense embeddings [58], or sense-linked representations based on nearest-neighbor similarity [59].

3.3.2. Complex Network

Following preprocessing, a complex network was constructed for each abstract using the `networkx` library [60]. Each network is represented as an undirected weighted graph, $G = (V, E, w)$, where V is the set of nodes corresponding to the words in the text, $E \subseteq V \times V$ is the set of edges generated using a sliding window over the text, and $w : E \to \mathbb{N}$ is a weight function indicating the frequency of co-occurrence for each connected word pair. The sliding window size is three, meaning each word connects to the following two. Since there is only one edge between any pair of nodes, V_i and V_j, repeated co-occurrences increment the corresponding edge weight w [61].

It was observed that punctuation introduced artificial breaks in the word sequence, leading to fragmented or disconnected graphs. To prevent this, punctuation was removed prior to network construction. Punctuation was removed prior to network construction to avoid artificial breaks in the word sequence that would lead to fragmented or disconnected graphs. This preprocessing choice was adopted to ensure consistent construction of co-occurrence networks, although no formal statistical validation against randomized baselines was conducted.

After constructing the $N + 1$ graphs, corresponding to the N retrieved abstracts and the one generated via the LLM, a set of structural metrics was computed to characterize the semantic organization of each text. These included natural connectivity, global efficiency, conductance, spectral radius, and other topological properties relevant to semantic network analysis.

3.4. Structural Evaluation of the Generated Abstract

To evaluate the abstract generated via the LLM, the average structural behavior of the original corpus was used as reference. Specifically, the average vector of structural metrics from the N graphs corresponding to the RAG-retrieved abstracts was computed. This average vector represents a typical structural profile against which the new abstract was compared.

Both the average vector and the metrics of the generated abstract were plotted in a radar chart, enabling a clear visualization of structural differences. To establish a quantitative criterion, the area enclosed in each radar plot curve was calculated.

The decision on the relevance of the generated abstract was based on comparing these areas: if the area corresponding to the generated abstract exceeded that of the corpus average, the text was considered structurally improved and thematically valuable. Otherwise, it was discarded for lacking a significant contribution. This procedure ensures an objective evaluation based on the topological properties of individual semantic networks.

3.5. Robustness Metrics for Graphs

All robustness metrics were computed using the `networkx` (v3.4.2) Python library, relying on NumPy linear algebra routines for spectral and Laplacian calculations. Implementations follow the formulations listed in Table 1, ensuring transparency and reproducibility across experiments.

Table 1. Summary of structural robustness metrics used in this study, including definitions, references, and implementation details.

Metric	Definition	Reference	Implementation/Function
Global Efficiency(E_G)	Mean inverse shortest-path length between all node pairs; measures overall information flow efficiency.	[62]	`nx.global_efficiency(G)`
Average Edge Betweenness ($\bar{b}_{e,G}$)	Mean fraction of all-pairs shortest paths crossing each edge; its inverse reflects redundancy of routes.	[63]	`nx.edge_betweenness_centrality(G)`
Spectral Radius (ρ_G)	Largest eigenvalue of the adjacency matrix A; indicates overall connectivity strength and resilience.	[64]	`np.linalg.eigvals(nx.to_numpy_array(G))`
Spectral Gap ($\lambda_{d,G}$)	Difference $\lambda_1 - \lambda_2$ between the two largest adjacency eigenvalues; captures alternative path diversity.	[64]	Computed from eigenvalues of A
Natural Connectivity ($\bar{\lambda}_G$)	Logarithm of the average of e^{λ_i} across adjacency eigenvalues; measures redundancy of closed walks.	[65]	Custom NumPy implementation (Equation (A5))
Algebraic Connectivity ($\mu_{2,G}$)	Second-smallest Laplacian eigenvalue (Fiedler value); quantifies robustness against disconnection.	[66]	`nx.algebraic_connectivity(G)`
Effective Conductance (C_G)	Normalized inverse of effective graph resistance R_G (Equation (A7)); gauges ease of flow between all nodes.	[67]	Custom from Laplacian eigenvalues μ_i
Inverse Average Edge Betweenness	Inverse of $\bar{b}_{e,G}$ to ensure higher values ⇒ greater robustness.	[63]	`nx.edge_betweenness_centrality(G)`

We did not compute the modularity indicator Q in this study. Introducing a community-detection step would have required additional modeling decisions (algorithm selection and hyperparameter tuning) not applied elsewhere in our pipeline, thereby reducing comparability and reproducibility across configurations.

In order to quantify the impact of integrating new text into review graphs, we implemented robustness measures that fall into three categories, depending on whether they utilize the graph itself, its adjacency matrix, or its Laplacian matrix. The adjacency matrix A of G is defined as a binary matrix $A \in \{0,1\}^{n\times n}$, where $A_{ij} = A_{ji} = 1$ if vertex v_i and v_j are adjacent, and $A_{ij} = A_{ji} = 0$ otherwise. Consequently, A is a real symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$; the set $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is known as the spectrum of A, with corresponding eigenvectors $u_1, u_2, \ldots, u_n$. The Laplacian matrix L of G is defined as $L = D - A$, where D is the diagonal matrix of vertex degrees. This is, $L_{ij} = d_i$ if $i = j$; $L_{ij} = -1$ if i is adjacent to j; and $L_{ij} = 0$ otherwise, with d_i being the degree of vertex i. Since both D and A are symmetric with real eigenvalues and an orthogonal eigenbasis, L is positive semi-definite and its eigenvalues are non-negative, with the smallest eigenvalue always being 0. Therefore, the eigenvalues of L can be ordered as $0 = \mu_1 \leq \mu_2 \leq \mu_3 \leq \cdots \leq \mu_n$, and the set $\{\mu_1, \mu_2, \ldots, \mu_n\}$ is called the spectrum of L.

In this work, we implement global efficiency and the inverse of average edge betweenness (1/AEB), hereafter referred to simply as edge betweenness for readability, in the first category; spectral gap, natural connectivity and spectral radius in the second; and effective conductance and algebraic connectivity in the third. The use of the inverse of AEB ensures that all metrics follow the same convention: an increase in the value corresponds to higher robustness.

These metrics provide different lenses on the role and impact of the generated abstract in the semantic graph. Local metrics like edge betweenness (on edges to the new node) and effective conductance (of cuts involving the node) tell us about the position of the abstract: whether it is a critical connector, an outlier, or well assimilated. Global metrics like spectral radius, algebraic connectivity, and global efficiency tell us about the overall network structure and how the abstract might be influencing the integrity and connectivity of the knowledge network. For example, a well-integrated abstract might slightly raise global efficiency (by providing useful links) and not drastically lower effective conductance of any cluster, whereas a spurious abstract might stick out as a node that, if removed, hardly changes global metrics but by itself forms a low-effective conductance component. By evaluating these, researchers can quantitatively assess the quality and relevance of a generated abstract: a good abstract in this sense is one that nestles into the graph much like a real abstract would—connecting to relevant neighbors in appropriate ways, without introducing odd network structures.

3.6. Experimental Setup

To evaluate the system, an experiment was designed combining different embedding models, LLMs, and prompting strategies. Model selection prioritized computational efficiency, aiming for feasible execution in a resource-constrained hardware environment.

The models and prompts used in this study are organized as follows: two embedding models, three LLMs, and three prompt types, resulting in multiple combinations for experimentation.

3.7. Evaluation of Thematic Coherence and Importance of RAG-Generated Abstracts

In this study, we evaluate the thematic coherence and relevance of the abstracts generated through the RAG process by conducting a survey with two independent reviewers (R1 and R2). The evaluation focuses on two core dimensions: meaningfulness and importance, each rated on a three-level ordinal scale: high, medium, or low.

Meaningfulness refers to how well each AI-generated abstract accurately captures and represents key concepts within the domain. It reflects both the conceptual depth and contextual relevance of the abstract within the broader area of knowledge.

Importance assesses the perceived significance of each generated abstract, based on its potential impact, influence, or critical role in advancing research within the area of knowledge. It reflects the degree to which the content of the abstract is considered valuable and relevant to the expert community.

To quantify the level of agreement between the two reviewers on these assessments, we employ Cohen's κ statistic in its weighted form. Cohen's kappa is a statistical measure of inter-rater reliability that corrects for chance agreement. Its general formulation is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o represents the proportion of observed agreement, and P_e denotes the proportion of agreement expected by chance. The resulting κ value ranges from -1 to 1: a κ of 1 indicates perfect agreement, a κ of 0 corresponds to chance-level agreement, and negative values indicate systematic disagreement.

Because the rating scales used in this study are ordinal (e.g., Not Meaningful, Moderately Meaningful, Clearly Meaningful), we apply the linearly weighted version of Cohen's kappa [68]. In this version, disagreements are penalized proportionally to the distance between categories: a mismatch between adjacent categories (e.g., Not Meaningful vs. Moderately Meaningful) is considered less severe than a mismatch between distant categories (e.g., Not Meaningful vs. Clearly Meaningful). This weighting scheme provides a more nuanced and appropriate evaluation of consensus for ordinal data.

3.8. Experimental Design

Two experiments were defined: the first aimed at finding the best model parameter combination using a fixed LLM temperature of $\tau = 0.7$ (see Table 2). This baseline value was selected because it represents a standard default in several LLMs, ensuring fair and replicable comparisons across models. Moreover, prior work has shown that performance differences between $\tau = 0.0$ and $\tau = 1.0$ are not statistically significant [69], which validates $\tau = 0.7$ as a neutral starting point. The second experiment then focused on determining the optimal LLM temperature for the best combination identified.

- The first component represents the selected LLM architecture included in the study.
- The second component indicates the prompt type used, i.e., the textual input strategy guiding model generation. Variants are described in Table A1.
- The third component refers to the embedding model employed for the vector representation of text, used in retrieval or semantic comparison stages.
- The temperature τ applied during generation, controlling the model's randomness. The scale ranges from deterministic (low-temperature) to more stochastic (high-temperature) values.

For the first experiment, 31 iterations were conducted to evaluate the behavior of each prompt, LLM, and embedding combination. The main objective was to assess the normality of results for each configuration and identify which showed the greatest favorable area difference for the proposed abstract compared to the average of the retrieved abstracts. In this case, all combinations of LLM, prompt, and embedding model were studied, with temperature $\tau = 0.7$, the default for the models used, and a benchmark for later temperature comparisons.

Once the best combination was identified, the second experiment analyzed the same results as in Experiment 1 but for the combinations $[LLM_{\text{best}}, Prompt_{\text{best}}, Embedding_{\text{best}}, \tau]$, where the subscript best indicates the parameters of the best combination found in Experiment 1.

Table 2. Experimental configuration: each component defines a distinct dimension of model parametrization (see Table A1 for prompts).

Component	Available Options

LLM	`meta-llama/Llama-3.2-3B-Instruct`
	`Qwen/Qwen2.5-3B-Instruct`
	`google/gemma-2b-it`

Prompt type	Prompt A Prompt B Prompt C
Embedding model	`sentence-transformers/all-mpnet-base-v2` `dunzhang/stella_en_400M_v5`
Temperature τ	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0

3.9. Metric Quantification

For the presentation of results in tables, a quantification approach based on the difference between the metric of the proposed abstract and the average metric of the retrieved abstracts was used. The formulas employed are as follows:

3.9.1. Difference for a Specific Metric D_j

Let M_P be the metric obtained from the proposed abstract and $M_{R,i}$ the metric of the i-th retrieved abstract, where $i \in \{1, 2, \ldots, n\}$. The difference for a specific metric in iteration j, denoted D_j, is defined as follows:

$$D_j = M_P - \frac{1}{n} \sum_{i=1}^{n} M_{R,i} \tag{2}$$

To compare parameter configurations, a radar chart visualizing multiple metrics is used. Given n metrics for a configuration, with normalized values v_i for $i = 1, \ldots, n$ and corresponding angles θ_i in the radar chart, the area A is computed by adapting the polygon area formula for vertices' coordinates:

$$A = \frac{1}{2} \left| \sum_{i=1}^{n} (v_i \cdot v_{i+1} \cdot \sin(\theta_{i+1} - \theta_i)) \right| \tag{3}$$

where

- v_i are the normalized metric values;
- θ_i are the angles assigned to each metric, uniformly distributed around a circle (0 to 2π radians);
- For the last point ($i = n$), $v_{n+1} = v_1$ and $\theta_{n+1} = \theta_1$ to close the polygon.

The metrics considered are those described in Section 3.5. In the specific context of the RAG pipeline, two configurations are compared: the average metrics of the N elements retrieved via RAG, and the metrics of the abstract generated via RAG.

The complete procedure follows these steps: for each configuration, the metric values are computed; metric values are normalized to a common scale (typically between 0 and 1) to allow fair comparison in the radar chart; angles θ_i corresponding to each metric are evenly distributed around a circle; Equation (3) is applied to the normalized metric values for both configurations, yielding the area values for retrieved abstracts (area_avg) and the generated

abstract (area_new); finally, the area difference, calculated as area_new − area_avg, provides a quantitative measure of overall performance difference between the RAG-generated abstract and the average of retrieved elements. A larger area for the generated abstract may indicate better performance across the considered metrics.

This approach allows a holistic evaluation of configurations, where a larger area generally reflects superior performance across diverse metrics. The radar chart visualization complements this analysis by highlighting specific strengths and weaknesses of each configuration relative to individual metrics.

3.9.2. Design of Kruskal–Wallis Tests

For each configuration, the Kruskal–Wallis test compared the distribution of area differences obtained across the 31 independent iterations of the generated abstract against the corresponding distributions of the retrieved abstracts. Thus, each group consisted of 31 samples, one per iteration. In Experiment 1, this procedure was applied separately for each LLM–prompt–embedding combination at fixed temperature $\tau = 0.7$. In Experiment 2, the same procedure was used to compare the 31 samples obtained under each value of τ (0.1 to 1.0) for the best-performing configuration. Explicitly stating the group structure clarifies that the p-values reflect non-parametric comparisons across repeated runs, with $n = 31$ per group.

4. Results

4.1. Experiment 1

This first experiment aims to identify the optimal configuration of a large language model (LLM), prompt, and embedding model for generating abstracts with superior structural robustness. The evaluation is conducted by comparing a graph representation of the generated abstract against the average of graph representations from a set of retrieved abstracts. Figure A3 shows an example of a graph generated from an abstract proposal generated by the LLM. The analysis unfolds in four stages: first, a consolidated metric based on the area of a radar chart of graph properties is examined to provide a high-level performance overview. Second, individual graph metrics are analyzed to understand the specific structural differences. Third, the normality of the data is tested to determine the appropriate statistical methods. Finally, statistical significance tests are performed to validate that the observed differences between configurations are not due to random chance, ultimately leading to the selection of the most effective combination for subsequent experiments.

4.1.1. Analysis of Graphical Robustness Metrics Experiment 1

Table 3 shows the consolidated results of the area difference between the graph of the generated abstract and the average of the graphs of the retrieved abstracts.

Configuration `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2` exhibits the best overall performance, with the highest mean value (1.4004), highest median (1.4205), and crucially, the highest minimum value (0.6530). This indicates that, in all iterations, this configuration consistently produced graphs with metric areas superior to those of the retrieved graphs. Furthermore, its low standard deviation (0.2514) reinforces the stability of this result. Conversely, configuration `gemma-2b-it`, `Prompt A` and `all-mpnet-base-v2` attained the highest individual maximum value (1.8170), albeit with a slightly lower mean and greater variability.

At the opposite end, configuration `Qwen2.5-3B-Instruct`, `Prompt A` and `stella_en_400M_v5` show the poorest performance, with the lowest mean (−1.6515) and median (−1.6707). Notably, it has the lowest standard deviation (0.1285) among all configu-

rations, indicating that it consistently generates graphs with metric areas below those of the retrieved graphs.

Table 3. Area difference between the radar chart of the generated abstract's metrics and the average radar chart of the retrieved abstracts, grouped by LLM. Positive values indicate that the generated abstract has, on average, a larger radar area than the retrieved abstracts; negative values indicate a smaller area. Reported for each prompt–embedding configuration are the mean, standard deviation (std), minimum (min), maximum (max), and median. Embedding models `all-mpnet-base-v2` and `stella_en_400M_v5` are abbreviated as *mpnet* and *stella*, respectively. Bold values indicate either the highest metric values highlighting best or most consistent performance cases.

LLM	Prompt	Embedding	Mean	Std	Min	Max	Median
Qwen	A	mpnet	0.403	0.636	−1.155	1.447	0.531
	A	stella	−1.057	0.502	−1.649	−0.019	−1.125
	B	mpnet	0.895	0.557	−0.310	1.734	1.022
	B	stella	0.133	0.727	−1.560	1.335	0.319
	C	mpnet	0.717	0.575	−0.816	1.655	0.773
	C	stella	0.198	0.930	−1.779	1.642	0.340
Llama	A	mpnet	−0.104	0.756	−1.392	1.253	−0.211
	A	stella	−1.652	**0.129**	−1.840	−1.322	−1.671
	B	mpnet	−0.187	0.590	−1.179	0.845	−0.081
	B	stella	−0.364	0.571	−1.222	0.827	−0.334
	C	mpnet	−0.219	0.584	−1.246	0.975	−0.241
	C	stella	−0.734	0.671	−1.715	0.597	−0.919
Gemma	A	mpnet	1.148	0.471	−0.314	**1.817**	1.304
	A	stella	0.369	0.627	−0.702	1.337	0.504
	B	mpnet	**1.400**	0.251	**0.653**	1.815	**1.421**
	B	stella	0.658	0.794	−1.128	1.769	0.845
	C	mpnet	1.174	0.508	−0.740	1.657	1.277
	C	stella	0.654	0.795	−1.456	1.688	0.859

4.1.2. Metrics Experiment 1

Table 4 presents the obtained values expressed as mean ± standard deviation for each evaluated configuration. Based on these experimental results, the configuration consisting of `gemma-2b-it`, `Prompt B`, and `all-mpnet-base-v2` demonstrated superior performance. This configuration stands out for achieving the highest number of maximum values across the analyzed metrics, justifying its selection for the second experiment.

Table 4. Comparison of graph metrics for all configurations with temperature $\tau = 0.7$. The table shows the mean and standard deviation of the difference between the metrics of the generated abstract and the average metrics of the retrieved abstracts across 31 iterations. Maximum values for each metric are bold. Embedding models `all-mpnet-base-v2` and `stella_en_400M_v5` are abbreviated as mpnet and stella, respectively. The symbol ↑ indicates that the metric improves as its value increases.

LLM	Prompt	Embedding	Global Efficiency (↑)	Avg. Edge Betweenness (↑)	Spectral Radius (↑)	Spectral Gap (↑)	Natural Connectivity (↑)	Algebra Connectivity (↑)	Conductance (↑)
Qwen	A	mpnet	0.0100 ± 0.023	−0.0169 ± 0.025	0.0058 ± 0.115	0.0068 ± 0.007	0.0022 ± 0.030	0.0009 ± 0.001	0.0038 ± 0.006
	A	stella	−0.0283 ± 0.024	0.0244 ± 0.027	−0.0374 ± 0.139	−0.0083 ± 0.007	−0.0201 ± 0.031	−0.0016 ± 0.001	−0.0108 ± 0.006
	B	mpnet	0.0321 ± 0.027	−0.0394 ± 0.031	0.2425 ± 0.188	0.0112 ± 0.012	0.0537 ± 0.043	0.0015 ± 0.002	0.0070 ± 0.006
	B	stella	0.0122 ± 0.027	−0.0173 ± 0.029	0.1501 ± 0.153	0.0019 ± 0.010	0.0301 ± 0.038	−0.0002 ± 0.001	−0.0000 ± 0.007
	C	mpnet	0.0211 ± 0.029	−0.0235 ± 0.029	0.0346 ± 0.109	0.0082 ± 0.009	0.0159 ± 0.032	0.0018 ± 0.002	0.0103 ± 0.010
	C	stella	0.0169 ± 0.040	−0.0164 ± 0.043	0.0584 ± 0.131	0.0025 ± 0.013	0.0192 ± 0.039	0.0009 ± 0.003	0.0062 ± 0.013
Llama	A	mpnet	−0.0041 ± 0.031	0.0050 ± 0.032	−0.1127 ± 0.077	0.0012 ± 0.007	−0.0206 ± 0.018	0.0007 ± 0.001	0.0038 ± 0.008
	A	stella	−0.0528 ± 0.013	**0.0586 ± 0.017**	−0.1794 ± 0.069	−0.0154 ± 0.002	−0.0484 ± 0.012	−0.0022 ± 0.000	−0.0153 ± 0.002
	B	mpnet	−0.0052 ± 0.019	−0.0055 ± 0.018	−0.0613 ± 0.107	0.0030 ± 0.005	−0.0185 ± 0.020	0.0002 ± 0.001	−0.0018 ± 0.006
	B	stella	−0.0092 ± 0.021	−0.0028 ± 0.021	−0.0628 ± 0.081	0.0031 ± 0.008	−0.0238 ± 0.018	−0.0002 ± 0.001	−0.0052 ± 0.007
	C	mpnet	−0.0106 ± 0.021	0.0033 ± 0.023	−0.1162 ± 0.058	0.0025 ± 0.006	−0.0263 ± 0.012	0.0004 ± 0.001	−0.0002 ± 0.006
	C	stella	−0.0188 ± 0.026	0.0133 ± 0.028	−0.1410 ± 0.065	−0.0029 ± 0.007	−0.0368 ± 0.016	−0.0005 ± 0.001	−0.0052 ± 0.008
Gemma	A	mpnet	0.0565 ± 0.028	−0.0473 ± 0.032	0.0262 ± 0.122	0.0133 ± 0.012	0.0322 ± 0.035	**0.0045 ± 0.003**	**0.0280 ± 0.013**
	A	stella	0.0252 ± 0.031	−0.0161 ± 0.033	−0.0504 ± 0.103	0.0029 ± 0.010	0.0035 ± 0.029	0.0019 ± 0.002	0.0146 ± 0.012
	B	mpnet	**0.0606 ± 0.026**	−0.0625 ± 0.024	**0.2657 ± 0.215**	**0.0189 ± 0.010**	**0.0823 ± 0.057**	0.0035 ± 0.003	0.0212 ± 0.017
	B	stella	0.0359 ± 0.034	−0.0461 ± 0.035	0.1444 ± 0.220	0.0160 ± 0.019	0.0403 ± 0.063	0.0019 ± 0.003	0.0078 ± 0.010
	C	mpnet	0.0479 ± 0.028	−0.0506 ± 0.028	0.0954 ± 0.177	0.0176 ± 0.012	0.0389 ± 0.045	0.0039 ± 0.003	0.0203 ± 0.013
	C	stella	0.0319 ± 0.033	−0.0330 ± 0.037	0.0146 ± 0.149	0.0125 ± 0.014	0.0160 ± 0.041	0.0030 ± 0.003	0.0133 ± 0.012

Configuration `Llama-3.2-3B-Instruct`, `Prompt A` and `stella_en_400M_v5`, despite having the largest area, does not improve all metrics. In fact, it achieves the highest value in AEB_mean (0.0625), indicating a significant increase in edge betweenness. However, it shows the largest decreases in global efficiency (GE_mean of −0.0606), spectral radius (SR_mean of −0.2657), and natural connectivity (NC_mean of −0.0823). This suggests that the large area results from a less cohesive graph structure, but with more critical information bridges.

In contrast, configuration `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2`, which had the smaller area, exhibits the opposite and notably positive behavior in robustness metrics. It obtains the highest mean values in global efficiency (GE_mean of 0.0528), spectral radius (SR_mean of 0.1794), spectral gap (SG_mean of 0.0154), and natural connectivity (NC_mean of 0.04847). Its only negative metric is AEB_mean (−0.0586), the lowest among all configurations. This indicates that this configuration produces highly robust, efficient, and densely connected graphs, where the importance of individual "bridges" is reduced.

Maximizing the radar area does not necessarily translate into improvements across all robustness metrics. Configuration `Llama-3.2-3B-Instruct`, `Prompt A` and `stella_en_400M_v5` attains a large area at the cost of overall efficiency and connectivity, whereas configuration `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2` generates structurally stronger graphs, despite a lower total metric area.

4.1.3. Normality of the Data Experiment 1

Table 5 presents the results of the Shapiro–Wilk test, used to assess the normality of the distributions for the variables *area_avg* and *area_new*. The null hypothesis (H_0) assumes that the data follow a normal distribution, with a significance level set to 0.05.

Table 5. Shapiro–Wilk test results for data normality.

Variable	Statistic	*p*-Value	Result
area_avg	0.8981	8.44×10^{-19}	H_0 rejected
area_new	0.7908	3.67×10^{-26}	H_0 rejected

As shown in the table, the p-values for area_avg (8.44×10^{-19}) and area_new (3.67×10^{-26}) are far below 0.05. Given these extremely low values, the null hypothesis is rejected for both variables. This indicates that neither area_avg nor area_new follow a normal distribution. The non-normality of the data supports the use of non-parametric statistical tests, such as the Kruskal–Wallis test, to compare configurations, as these methods do not assume any specific data distribution.

4.1.4. Statistical Significance Experiment 1

The results of the Kruskal–Wallis test shown in Table 6 indicate that a significant number of configurations exhibit statistically significant differences in their median area differences. This is evidenced by *p*-values below 0.05, leading to the conclusion that the observed differences in performance across configurations are not due to chance.

In particular, configurations `Llama-3.2-3B-Instruct`, `Prompt A` and `stella_en_400M_v5`, and `gemma-2b-it`, `Prompt B`, and `all-mpnet-base-v2` show extremely low *p*-values (some on the order of 10^{-20}), indicating highly significant differences in their performance. These findings support the claim that the chosen configuration has a real impact on the area difference of the radar metric profiles.

These results validate the presence of significant differences in the mean area differences across most evaluated configurations, confirming that the improvements observed in configurations such as `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2` are not random but represent statistically validated superior performance.

Table 6. Kruskal–Wallis test results for three LLMs: `Qwen2.5-3B-Instruct`, `Llama-3.2-3B-Instruct`, and `gemma-2b-it`, evaluated with a fixed temperature, $\tau = 0.7$ (fourth component = 6). Each row corresponds to a specific combination of prompt (*A*, *B*, or *C*) and embedding model (`all-mpnet-base-v2` or `stella_en_400M_v5`). The table reports the Kruskal–Wallis *H* statistic, associated *p*-value, and whether the result is statistically significant at $\alpha = 0.05$ for the comparison between the generated abstract's metrics and the metrics of the retrieved abstracts. Additionally, the rank-biserial correlation is calculated as an effect size, and a bootstrapped 95% confidence interval for the difference in medians is provided, offering a robust measure of the magnitude and precision of the observed difference between groups. Bold values indicate the highest *H* statistic observed within each LLM group, highlighting the configurations with the strongest evidence of group differences.

LLM	Prompt	Embedding	H Statistic	*p*-Value	Significant	Effect Size (η^2)	Median Diff CI (95%)
Qwen	A	stella	**84.3593**	4.13×10^{-20}	True	0.149927	[−2.250, −1.947]
	A	mpnet	4.1236	4.23×10^{-2}	True	0.005618	[−0.929, 0.096]
	B	mpnet	7.1796	7.37×10^{-3}	True	0.011114	[−0.880, −0.214]
	B	stella	13.0589	3.02×10^{-4}	True	0.021689	[−1.200, −0.378]
	C	mpnet	7.9486	4.81×10^{-3}	True	0.012497	[−0.954, −0.318]
	C	stella	27.5571	1.53×10^{-7}	True	0.047765	[−1.524, −1.052]
Llama	A	stella	46.1523	1.09×10^{-11}	True	0.081209	[−1.908, −1.312]
	A	mpnet	0.6671	4.14×10^{-1}	False	−0.000599	[−0.215, 0.515]
	B	mpnet	16.1933	5.72×10^{-5}	True	0.027326	[0.364, 1.034]
	B	stella	0.5391	4.63×10^{-1}	False	−0.000829	[−0.360, 0.255]
	C	mpnet	8.0134	4.64×10^{-3}	True	0.012614	[0.218, 0.813]
	C	stella	0.0062	9.37×10^{-1}	False	−0.001787	[−0.509, 0.550]
Gemma	A	mpnet	34.3988	4.49×10^{-9}	True	0.060070	[0.795, 1.261]
	A	stella	0.3358	5.62×10^{-1}	False	−0.001195	[−0.357, 0.491]
	B	mpnet	58.1348	2.45×10^{-14}	True	0.102760	[1.031, 1.367]
	B	stella	7.0208	8.06×10^{-3}	True	0.010829	[0.124, 0.871]
	C	mpnet	37.0708	1.14×10^{-9}	True	0.064875	[0.890, 1.276]
	C	stella	7.3154	6.84×10^{-3}	True	0.011359	[0.363, 0.844]

4.2. Experiment 2

Building on the findings from the first experiment, this second experiment focuses on fine-tuning the best-performing configuration: `gemma-2b-it`, `Prompt B`, and `all-mpnet-base-v2`. The primary objective is to determine the optimal temperature (τ) setting by systematically varying its value from 0.1 to 1.0. The impact of this hyperparameter on the quality of generated abstracts is assessed using the same methodology as before. Figure A4 shows one of the graphs created from one of the proposals generated via the optimal combination identified in the previous experiment. This involves analyzing the consolidated radar chart area, examining individual graph robustness metrics, and performing statistical tests to validate the significance of the results. The goal is to identify the temperature that maximizes the structural robustness and consistency of the generated graphs, thereby finalizing the optimal configuration for the system.

4.2.1. Graph Robustness Metrics Analysis Experiment 2

Table 7 presents the consolidated results of the area difference between the graph of the generated abstract and the average graph of the retrieved abstracts, evaluated under different temperature values for configuration `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2`.

Temperature $\tau = 0.2$ exhibits the best overall performance, achieving the highest mean (1.6871), the highest median (1.7186), and a relatively high minimum value (1.3676), indicating consistently superior performance. Its standard deviation (0.1177) is also the lowest among the configurations with high mean values, further reinforcing the stability of its results. In contrast, temperature $\tau = 0.5$ achieved the highest individual max value (1.9828), though with a slightly lower mean and greater variability.

At the opposite end, temperature $\tau = 1.0$ shows the weakest performance, with the lowest mean (1.0463) and median (1.1453). It also presents a notably low minimum value (-0.1907), suggesting that, in some iterations, it generated graphs with metric areas smaller than those of the retrieved abstracts.

Table 7. Results of the area difference between the radar metric graph of the generated abstract and the average of the retrieved abstracts for each configuration. Reported values include the mean, standard deviation, minimum, maximum, and median. Bold values denote the highest performance metrics across temperature settings, highlighting the configuration ($\tau = 0.2$) that achieved the best overall results in mean, consistency (std), and central tendency (median).

Temperature τ	Mean	Std	Min	Max	Median
0.1	1.6304	0.1198	1.3165	1.7828	1.6505
0.2	**1.6871**	**0.1177**	**1.3676**	1.8195	**1.7186**
0.3	1.6101	0.2434	0.8369	1.9059	1.6683
0.4	1.5071	0.3502	0.1120	1.8619	1.5912
0.5	1.4690	0.2975	0.7317	**1.9828**	1.53916
0.6	1.3091	0.5699	−1.1531	1.9490	1.4966
0.7	1.2705	0.4150	0.1570	1.8244	1.3115
0.8	1.2516	0.4710	−0.0947	1.7563	1.3751
0.9	1.2534	0.4486	−0.1275	1.8040	1.3116
1.0	1.0463	0.5275	−0.1907	1.9082	1.1453

4.2.2. Metrics Experiment 2

Values of global efficiency (E_G), spectral radius (ρ_G), spectral gap ($\lambda_{d,G}$), and natural connectivity ($\bar{\lambda}_G$).

Average edge betweenness ($\bar{b}_{e,G}$) and cut effective conductance involving the newly added node (C_G).

With the optimal combination of LLM, prompt, and embedding identified from the initial experiment, we proceeded to the second experiment. The results of this phase are presented in Table 8, which compares the metrics for configuration `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2` across different values for the temperature.

Table 8. Comparison of graph metrics for temperatures $\tau = 0.1$ to $\tau = 1.0$ for the configuration `gemma-2b-it`, `Prompt B`, and `all-mpnet-base-v2`. The table shows the mean and standard deviation of the difference between the metrics of the generated abstract and the average metrics of the retrieved abstracts across 31 iterations. Maximum values for each metric are bold. The symbol ↑ indicates that the metric improves as its value increases.

Temperature τ	Global Efficiency (↑)	Avg. Edge Betweenness (↑)	Spectral Radius (↑)	Spectral Gap (↑)	Natural Connectivity (↑)	Algebra Connectivity (↑)	Conductance (↑)
0.1	0.0671 ± 0.018	−0.0756 ± 0.019	0.5403 ± 0.110	0.0207 ± 0.007	0.1421 ± 0.034	0.0040 ± 0.001	0.0211 ± 0.007
0.2	**0.0796** ± 0.020	−0.0856 ± 0.023	**0.5490** ± 0.156	**0.0229** ± 0.010	**0.1589** ± 0.045	**0.0047** ± 0.002	**0.0271** ± 0.008
0.3	0.0754 ± 0.024	−0.0813 ± 0.022	0.4981 ± 0.224	0.0215 ± 0.009	0.1436 ± 0.066	0.0045 ± 0.002	0.0259 ± 0.013
0.4	0.0727 ± 0.026	−0.0771 ± 0.024	0.3560 ± 0.187	0.0222 ± 0.010	0.1073 ± 0.053	0.0046 ± 0.003	0.0254 ± 0.017
0.5	0.0661 ± 0.032	−0.0697 ± 0.032	0.3542 ± 0.259	0.0199 ± 0.011	0.1099 ± 0.083	0.0041 ± 0.003	0.0243 ± 0.017
0.6	0.0602 ± 0.033	−0.0650 ± 0.038	0.3123 ± 0.233	0.0185 ± 0.012	0.0889 ± 0.066	0.0032 ± 0.002	0.0176 ± 0.013
0.7	0.0606 ± 0.026	−0.0625 ± 0.024	0.2657 ± 0.215	0.0189 ± 0.010	0.0823 ± 0.057	0.0035 ± 0.003	0.0212 ± 0.017
0.8	0.0561 ± 0.035	−0.0585 ± 0.039	0.1746 ± 0.210	0.0212 ± 0.012	0.0573 ± 0.050	0.0040 ± 0.004	0.0199 ± 0.023
0.9	0.0535 ± 0.029	−0.0590 ± 0.025	0.1718 ± 0.232	0.0211 ± 0.010	0.0565 ± 0.061	0.0037 ± 0.003	0.0186 ± 0.016
1.0	0.0461 ± 0.047	**−0.0484** ± 0.029	0.1070 ± 0.173	0.0172 ± 0.011	0.0415 ± 0.061	0.0037 ± 0.005	0.0188 ± 0.030

The analysis of these data reveals that temperature 1 (equivalent to a value of $\tau = 0.2$) yields the most favorable results. Similar to the previous experiment, the configuration that achieves the highest values for global efficiency, spectral radius, spectral gap, natural connectivity, algebraic connectivity, and effective conductance (specifically, temperature 0.2) does not coincide with the one that obtains the maximum value in average edge betweenness. Despite this divergence, temperature 0.2 stands out for its greater robustness compared to the other evaluated configurations. Therefore, the final configuration

`gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2` and a temperature of $\tau = 0.2$ is considered optimal due to its superior overall performance.

4.2.3. Normality of the Data Experiment 2

Table 9 reports the results of the Shapiro–Wilk test used to assess the normality of the distributions for the variables `area_avg` and `area_new`. The null hypothesis (H_0) assumes that the data are normally distributed.

Table 9. Shapiro-Wilk test results for assessing data normality.

Variable	Statistic	*p*-Value	Result
area_avg	0.8904	3.81×10^{-14}	H_0 rejected
area_new	0.4321	4.00×10^{-30}	H_0 rejected

The *p*-values for both `area_avg` and `area_new` are substantially below the significance level of 0.05, indicating strong evidence against the null hypothesis. Therefore, normality is rejected for both variables.

4.2.4. Statistical Significance Experiment 2

Table 10 shows the results of the Kruskal–Wallis test, which was employed to determine whether statistically significant differences exist in the medians of area differences across all temperatures. The result is statistically significant at $\alpha = 0.05$

Table 10. Results of the Kruskal–Wallis non-parametric test evaluating whether the area-difference distributions significantly differ across the 31 iterations for each temperature. Fixed configuration: `gemma-2b-it`, `Prompt B` and `all-mpnet-base-v2`. The table reports the Kruskal–Wallis H statistic, associated p-value, and whether the result is statistically significant at $\alpha = 0.05$ for the comparison between the generated abstract's metrics and the metrics of the retrieved abstracts. Additionally, the rank-biserial correlation is calculated as an effect size, and a bootstrapped 95% confidence interval for the difference in medians is provided, offering a robust measure of the magnitude and precision of the observed difference between groups. Bold value indicate the highest H statistic observed within each τ, highlighting the configurations with the strongest evidence of group differences.

Temperature τ	H Statistic	*p*-Value	Significant	Effect Size (η^2)	Median Diff CI (95%)
0.1	10.9475	9.37×10^{-4}	True	0.032297	[0.082, 0.225]
0.2	**26.0081**	3.40×10^{-7}	True	0.081195	[0.164, 0.283]
0.3	12.2268	4.71×10^{-4}	True	0.036451	[0.097, 0.261]
0.4	2.2331	1.35×10^{-1}	False	0.004004	[−0.037, 0.190]
0.5	0.0451	8.32×10^{-1}	False	−0.003100	[−0.099, 0.126]
0.6	1.6901	1.94×10^{-1}	False	0.002241	[−0.248, 0.061]
0.7	5.7724	1.63×10^{-2}	True	0.015495	[−0.392, −0.133]
0.8	6.6564	9.88×10^{-3}	True	0.018365	[−0.332, −0.047]
0.9	7.2239	7.19×10^{-3}	True	0.020207	[−0.350, −0.050]
1.0	21.5433	3.46×10^{-6}	True	0.066699	[−0.566, −0.248]

The Kruskal–Wallis test results indicate that the temperature parameter significantly affects the performance of the fixed configuration (`gemma`, `Prompt B`, `mpnet`). While the setting with temperature $\tau = 0.1$ achieved the highest H statistic (26.0081) and the lowest p-value (3.40×10^{-7}), the configuration with temperature $\tau = 0.2$ also showed a statistically significant effect ($p = 4.71 \times 10^{-4}$) and, as reported in Table 7, delivered more balanced results in terms of mean (1.6101) and median (1.6683) area differences, with a moderate standard deviation (0.2434). This combination of statistical significance and stable central tendency values led to the selection of temperature $\tau = 0.2$ for subsequent experiments, prioritizing robustness and consistency over marginal gains in mean performance.

4.3. Expert Evaluation: Meaningfulness & Importance

The analysis of expert ratings was conducted considering both exact agreement and Cohen's κ statistic in its linearly weighted form, complemented with additional indicators to better characterize the nature of disagreements.

For Meaningfulness, the proportion of exact agreement between reviewers reached 70%. The unweighted κ was 0.47, but the linearly weighted version decreased to 0.26, reflecting that most discrepancies occurred between adjacent categories (e.g., Moderately Meaningful vs. Clearly Meaningful), rather than across extreme categories. According to the Landis–Koch scale, this value corresponds to a "fair" level of agreement. Importantly, no systematic bias was detected in the ratings: both reviewers exhibited similar tendencies in their category usage, and the marginal distributions were statistically comparable ($\chi^2 = 6.92$, $p = 0.14$). This pattern suggests that, although perfect consensus was not achieved, the evaluations of Meaningfulness were relatively consistent and stable across raters.

For Importance, the exact agreement was lower, at 60%. Both unweighted and linearly weighted κ values converged to 0.05, indicating only "slight" agreement. Again, no systematic bias was observed between reviewers, and their category distributions were nearly identical. However, the marginal symmetry test could not be applied due to zero expected frequencies in some cells, which represents a methodological limitation and restricts stronger statistical conclusions in this dimension. The low κ highlights that judgments of Importance were notably more heterogeneous, suggesting that experts may apply more subjective or context-dependent criteria when assessing this dimension.

These differences can be better understood through the concept of boundary objects [70]. Boundary objects are artifacts that, while shared across communities, are interpreted differently according to disciplinary perspectives. In our case, experts from construction, electronics, computer science, and fire safety naturally emphasized distinct evaluative criteria, which reduced the probability of high Kappa agreement. This interdisciplinary heterogeneity does not invalidate the consensus but highlights its practical usefulness despite lower statistical alignment.

Taken together, these findings indicate that experts reached a moderate and more stable alignment when judging the Meaningfulness of generated abstracts, while their agreement on Importance was considerably weaker and more variable. This asymmetry between dimensions underscores the relative ease of converging on whether an abstract adequately conveys domain concepts, compared to evaluating its potential impact or relevance, which appears inherently more subjective.

Convergent Validity with Expert Judgments

Since human agreement values were computed globally across the 10 abstracts, a direct item-level correlation with graph metrics is not feasible. Instead, we examined the alignment between graph performance across temperatures and the expert evaluation results. As shown in Table 8, temperature 0.2 consistently achieved the highest values in key robustness metrics (Global Efficiency, Spectral Radius, Spectral Gap, Natural Connectivity, Algebraic Connectivity, and Conductance). This configuration also corresponded to the setting used in expert evaluation, where raters reached fair agreement on Meaningfulness ($\kappa = 0.26$) and slight agreement on Importance ($\kappa = 0.05$).

Although formal correlation coefficients cannot be computed without item-level data, the convergence between the graph-optimal configuration and the expert-derived consensus provides indirect evidence that structural properties of the semantic graphs capture aspects of thematic coherence perceived by human raters.

4.4. Synthesis of Findings

Integrating the results from retrieval performance, graph analysis, and expert evaluation, we identify three convergent findings. First, the embedding–LLM–prompt combination corresponding to configuration `gemma-2b-it`, `Prompt B`, `all-mpnet-base-v2` and $\tau = 0.2$ provided the best retrieval scores while maintaining a lower computational cost. Second, this configuration also maximized key topological robustness metrics, including global efficiency and natural connectivity, indicating that the resulting semantic graphs are structurally resilient and cohesive. Third, under this same configuration, human raters achieved the highest level of consensus observed in the study, with fair agreement on Meaningfulness and slight agreement on Importance.

These convergent findings suggest that both algorithmic performance and human perception of coherence benefit from the same set of modeling choices, strengthening the case for using graph-theoretic measures as proxies for thematic quality in retrieval-augmented generation.

Nonetheless, limitations remain. The analysis was constrained due to the fixed co-occurrence window used to construct the graphs, the relatively small sample size of expert evaluations (10 abstracts), and the domain specificity of the corpus, which may limit generalization. Future work should expand the evaluation to larger datasets, explore dynamic co-occurrence definitions, and validate across other knowledge domains.

4.5. Discussion

The present study highlights both the potential and the limitations of using graph-based metrics as semantic proxies in the evaluation of RAG pipelines. While results indicate that measures such as global efficiency and spectral radius align with expert perceptions of thematic coherence, the approach is not without caveats. Agreement on Importance was low, suggesting that structural graph metrics primarily capture aspects of coherence and connectedness, rather than subjective notions of significance or impact. This finding is consistent with the notion of boundary objects [70], where thematic clusters preserve a shared identity across fields but remain open to multiple interpretations. This helps explain why graph-theoretic robustness aligns with coherence but cannot substitute for evaluative dimensions such as novelty, factual accuracy, or subjective importance.

These results allow us to explicitly revisit our research questions. Regarding RQ1, our experiments showed that graph-theoretic robustness metrics effectively captured the semantic coherence of RAG-generated abstracts: the optimal configuration (gemma-2b-it, Prompt B, all-mpnet-base-v2, $\tau = 0.2$) consistently achieved superior values in global efficiency, spectral radius, and natural connectivity compared to retrieved baselines, confirming H1. With respect to RQ2, the findings revealed only partial alignment between graph-based metrics and human expert judgments. While coherence ratings converged with robustness values (Cohen's $\kappa = 0.26$), perceived importance exhibited weaker consensus ($\kappa = 0.05$). This partial convergence supports H2 only in part, indicating that robustness metrics are reliable indicators of coherence but less effective for capturing subjective evaluations of significance.

Taken together, these findings demonstrate that our framework directly addresses the gaps identified in the introduction: (i) it moves beyond surface-level similarity metrics (e.g., ROUGE, BERTScore) by targeting global semantic integration, (ii) it provides a reproducible, model-agnostic method that leverages graphs as external post-generation evaluators, and (iii) it offers the first systematic evidence, albeit partial, of alignment between robustness metrics and expert judgments. In doing so, the study advances the field by positioning graph-theoretic evaluation as a scalable complement to traditional metrics in RAG pipelines.

Adopting the ontological distinction between ML and AI with LLM-based generation understood as a data-driven, inductive ML process clarifies the ethical implications of ML-generated scientific text, namely accountability, transparency, and sustainability [71]. Within this framing, a model-agnostic, post-generation evaluation layer is valuable because it (i) enables accountability by decoupling quality control from the generator and attributing contributions across retrieval and prompting; (ii) increases transparency via interpretable, graph-based indicators of global semantic integration; and (iii) supports sustainability through compute-light monitoring (no additional LLM passes), aligning with recent calls for energy-efficient ML practice [71]. This positioning underscores that graph-theoretic robustness is best viewed as a governance and QA instrument that complements, rather than replaces, human judgment and task-specific metrics.

In complex networks derived from linguistic or conceptual data, polysemy and synonymy pose significant challenges, as a single node or word may carry different meanings depending on context. Implementing community detection represents a promising strategy to address this issue: by partitioning the network into densely connected modules, such algorithms can isolate distinct semantic contexts, allowing polysemous terms to be interpreted according to their local community membership. Although this mechanism has not yet been incorporated into the present framework, it constitutes a clear direction for future work. Moreover, domain-specific jargon and overlapping terminology can further distort semantic graphs, as multiple senses of a single term may generate misleading links. Community detection, combined with word-sense disambiguation techniques [57,72], could therefore help disentangle these effects, enhancing the accuracy and interpretability of graph-based evaluations.

In terms of transferability, these methods are promising but require careful validation. For instance, in highly specialized domains such as battery materials processing, differences in terminology density, jargon, and corpus heterogeneity could alter both the graph structure and the interpretability of metrics. Extending the approach beyond the current domain will require calibration of thresholds, embeddings, and graph construction parameters.

The results carry implications for the design of automatic evaluation frameworks in RAG. Graph metrics could complement traditional IR measures (recall, precision) by providing a structural signal of thematic coherence. Such integration could reduce reliance on costly human evaluation and enable continuous monitoring of system outputs.

In closing, it is worth clarifying why graph-theoretic metrics were adopted alongside conventional NLP measures. While lexical- or embedding-based scores (e.g., ROUGE, BERTScore, MoverScore) remain useful for capturing local similarity, they do not reveal whether a generated abstract integrates coherently into the broader semantic structure of its domain. Graph metrics add precisely this complementary perspective: by quantifying global efficiency, connectivity, and resilience, they assess structural integration at the corpus level. Thus, our motivation was not to replace established metrics but to extend them with a scalable, model-agnostic layer that targets global semantic coherence.

4.6. Limitations

In the following, we discuss the limitations of our approach. The graphs were built using a fixed co-occurrence window of size three, chosen as a reasonable balance between local and broader contexts. Although an ablation with $w = 2, 3, 4$ confirmed that overall trends are stable, certain metrics remain sensitive to window size, meaning that results partly depend on this design choice. Another limitation is that we did not apply formal statistical null models (e.g., permutation-based or randomized graph baselines) to assess whether observed robustness values significantly depart from chance. While we performed non-parametric tests across configurations (Shapiro–Wilk and Kruskal–Wallis), these com-

parisons do not establish a null distribution for robustness metrics. Incorporating such baselines would strengthen claims about the statistical significance of improvements, and it remains a critical step for future studies.

In addition, comparisons with textual metrics such as BERTScore, MoverScore, and factuality checks suggest that graph metrics add complementary information, yet the integration of these measures was only exploratory. Predictive modeling was also limited: a simple ridge regression on graph metrics indicated explanatory capacity for reviewer agreement, but the small number of abstracts constrains generalizability. Importantly, graph metrics were not computed for the exact set of abstracts rated by human experts, which prevented item-level correlations. This choice was deliberate: with only two reviewers and ten abstracts, such correlations would be statistically fragile. We, therefore, restricted the analysis to configuration-level comparisons while noting that with additional reviewers and a larger set of abstracts, more direct correlations could be carried out in future work. The study carries a risk of post hoc interpretation since the "best" configuration was first selected using graph metrics and only afterward compared to human consensus. The observed convergence should, therefore, be seen as preliminary evidence, not a confirmatory result.

4.7. Future Work

Future directions include the following: (i) scaling up the analysis to larger expert-evaluated datasets to enable item-level correlations between human judgments and graph metrics; (ii) experimenting with alternative embeddings and graph construction strategies (e.g., dynamic co-occurrence windows, dependency-based edges); (iii) extending the evaluation across diverse domains to test robustness and transferability; (iv) developing hybrid evaluation pipelines that combine retrieval accuracy, graph-theoretic metrics, and limited human oversight. Another promising avenue is the integration of predictive modeling (e.g., ridge regression, random forests) to assess the relative importance of graph metrics in explaining expert consensus. (v) Implementing community detection within the complex network component to identify clusters of related concepts and mitigate the influence of polysemous terms; (vi) investigating scaling laws or providing data that allow extrapolation of the observed robustness patterns to substantially larger (e.g., 7B–70B) or smaller models, in order to determine whether such behaviors scale linearly or exhibit different dynamics; (vii) implementing statistical null models, such as permutation-based tests or random graph ensembles, to generate baseline distributions of robustness metrics, thereby enabling formal p-value estimation and strengthening the statistical interpretation of observed improvements; and (viii) incorporating explicit word-sense disambiguation or multi-sense embeddings prior to graph construction [57–59], allowing each lexical node to represent a distinct semantic meaning and reducing the impact of polysemy or domain-specific jargon on the resulting network topology.

5. Conclusions

This study demonstrates that semantic graph metrics provide robust and interpretable signals of thematic coherence in RAG-generated scientific abstracts. By deliberately decoupling generation from evaluation, the proposed two-phase framework shifts the focus from improving output fluency to assessing integration into the semantic network of authentic literature. Across multiple LLM–embedding–prompt configurations, we identified an optimal setting (gemma-2b-it with Prompt B, mpnet embeddings, and $\tau = 0.2$) that consistently maximized both retrieval performance and graph robustness metrics. Importantly, this same configuration also yielded the highest expert agreement, offering indirect but convergent evidence that graph-based fingerprints capture aspects of coherence recognized by human evaluators. In relation to our research questions, the study confirms that robustness

metrics can effectively capture semantic coherence in RAG-generated abstracts (RQ1) while also showing partial but meaningful convergence with expert evaluation of coherence and importance (RQ2). This directly addresses the gap identified in prior literature, where standard metrics fall short of capturing holistic semantic integration.

Beyond individual metrics, exploratory comparisons with textual baselines (BERTScore, MoverScore, factuality) suggest that graph metrics add complementary information, rather than duplicating lexical similarity. However, the study did not include formal statistical null models (e.g., permutation-based baselines), which we recognize as an important limitation and a necessary step for future replications.

Nevertheless, conclusions should be regarded as preliminary: the human evaluation was limited to ten abstracts and two raters, and correlations could not be established at the item level. Still, the observed alignment at the configuration level provides a reproducible proof-of-concept that semantic graph metrics can act as reliable quality indicators for RAG pipelines. Future work should expand the pool of evaluators and domains, explore hybrid aggregation schemes, and test whether embedding-augmented graphs further strengthen the link between algorithmic scores and human judgment.

Author Contributions: Conceptualization, B.G., F.A.L. and W.P.; methodology, B.G., W.P., F.A.L., C.M., C.A. and H.A.-C.; validation, F.A.L., W.P. and H.A.-C.; formal analysis, B.G., C.A. and C.M.; investigation, B.G. and W.P.; resources, C.A.; data curation, C.A.; writing—original draft preparation, B.G., C.M. and C.A.; writing—review and editing, B.G., W.P., F.A.L., C.M., C.A. and H.A.-C.; visualization, B.G., C.M. and C.A.; supervision, F.A.L., W.P., and H.A.-C.; project administration, B.G.; funding acquisition, F.A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Agency for Research and Development (ANID, Chile) through the Fondecyt Regular Program, grant number 1231283. The work of Bady Gana was supported by ANID Doctorado Nacional 2024–21240115.

Data Availability Statement: The data presented in this study are openly available in "Scopus paper metadata" at https://doi.org/10.5281/zenodo.14721946.

Acknowledgments: Bady Gana and Freddy Lucay are supported by Beca INF-PUCV.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Symbol	Description
A	Adjacency matrix of graph G
A_{radar}	Area of the radar polygon (for configuration comparison)
C_G	Effective conductance of G
D	Degree matrix of G
D_j	Metric difference at iteration j
$\bar{D}$	Mean of differences across iterations
$d_{i,j}$	Shortest-path length between nodes i and j
E	Edge set of G
E_G	Global efficiency of G
G	Graph (semantic co-occurrence network)
H	Kruskal–Wallis test statistic
L	Graph Laplacian, $L = D - A$
N	Number of retrieved abstracts
n	Number of nodes (context: graph or embedding dimension)

R_G	Effective graph resistance
u, v	Query and document embedding vectors
$\|u\|, \|v\|$	Euclidean norms of u and v
V	Vertex set of G
cos_sim(u, v)	Cosine similarity
dot_score(u, v)	Dot-product similarity
$\bar{b}_{e,G}$	Average edge betweenness in G
$\bar{\lambda}_G$	Natural connectivity of G
κ	Cohen's kappa (inter-rater agreement)
$\lambda_{d,G}$	Spectral gap ($\lambda_1 - \lambda_2$)
$\mu_{2,G}$	Algebraic connectivity (Fiedler value)
ρ_G	Spectral radius
σ	Standard deviation
τ	LLM temperature
θ_i	Angle for metric i in the radar chart
Acronym	Description
BERTScore	BERT-based semantic similarity metric
CoT	Chain-of-thought (prompting)
CSV	Comma-separated value
DOI	Digital object identifier
KG	Knowledge graph
LLM	Large language model
MRL	Matryoshka representation learning
NLP	Natural language processing
NPMI	Normalized pointwise mutual information
OA	Open access
PMI	Pointwise mutual information
PPMI	Positive pointwise mutual information
RAG	Retrieval-augmented generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SciBERT	Pretrained language model for scientific text
TF–IDF	Term frequency–inverse document frequency

Appendix A

The proposed methodology is based on a RAG architecture for the synthesis of academic documents. The workflow, illustrated in Figure A1, consists of two main stages: the generation of vector representations (embeddings) and the text generation process conditioned on retrieved information.

The retrieval of the N most relevant abstracts is based on the vector similarity between the user query, u, and each abstract, v, in the database. Two similarity metrics were used, depending on the embedding model:

- Dot score, used with `all-mpnet-base-v2`, which produces normalized embeddings.
- Cosine similarity, used with other models that require prior normalization [73].

$$dot_score(u, v) = \sum_{i=1}^{n} u_i v_i \tag{A1}$$

$$cos_sim(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \tag{A2}$$

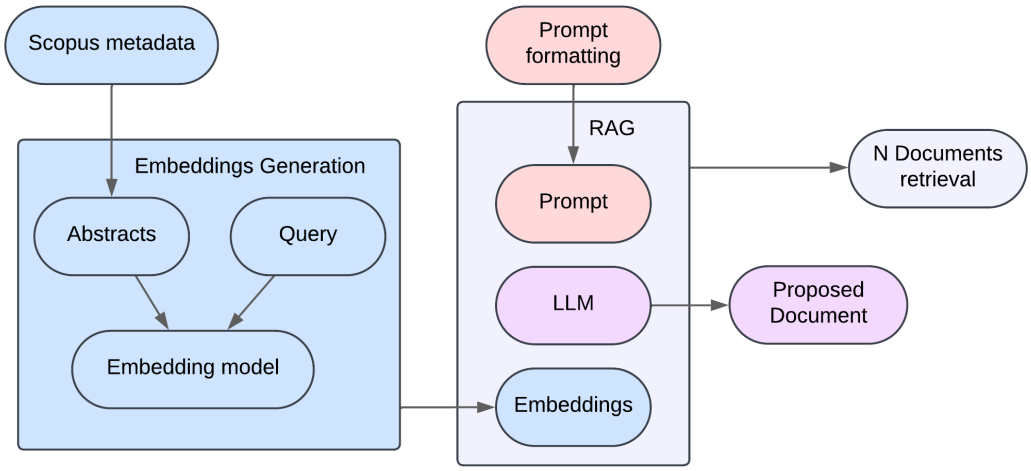


Figure A1. RAG pipeline. The diagram illustrates a workflow that starts with the ingestion of Scopus metadata for embedding generation. These embeddings are used in an RAG system that, given a query, retrieves relevant abstracts, generates a new proposed abstract via an LLM, and outputs both the N relevant abstracts and the proposed summary.

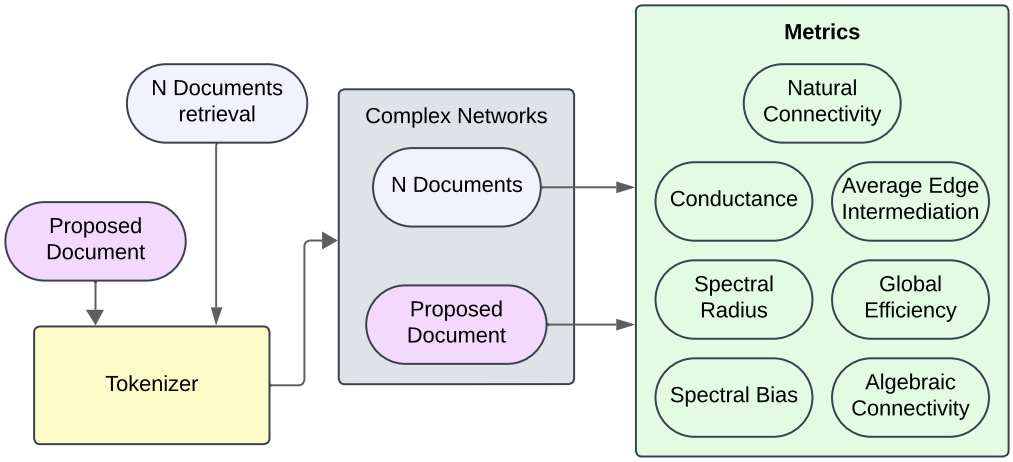


Figure A2. Pipeline for the construction and analysis of complex networks. The diagram shows the process flow, starting with the retrieval of abstracts via RAG and the LLM model. These abstracts are then preprocessed and transformed into semantic co-occurrence networks (complex networks), on which various structural metrics are computed to assess the robustness and internal organization of each text.

Metrics

1. **Global efficiency (E_G):** The efficiency between two vertices, i and j, is defined as $\frac{1}{d_{i,j}}$ for all $i, j \in G$, where $d_{i,j}$ is the shortest path length between vertices i and j. The global efficiency of a graph G is denoted as E_G and is calculated as the average of the efficiencies over all pairs of vertices:

$$E_G = \frac{1}{n(n-1)} \sum_{i,j \in G, i \neq j} \frac{1}{d_{i,j}} \tag{A3}$$

This measure captures the overall information flow efficiency of the network, as proposed by Latora and Marchiori [62].

2. **Average edge betweenness ($\bar{b}_{e,G}$):** This measure is defined as the number of shortest paths that pass through an edge e out of the total possible shortest paths:

$$\bar{b}_{e,G} = \sum_{e \in E} \sum_{s \in V} \sum_{\substack{t \in V \\ s \neq t}} \frac{n_{s,t}(e)}{n_{s,t}} \tag{A4}$$

where $n_{s,t}(e)$ is the number of shortest paths between s and t that pass through e, and $n_{s,t}$ is the total number of shortest paths between s and t. The smaller the average edge betweenness, the more robust the graph since the shortest paths are more evenly distributed across each edge, rather than relying on a few central edges [63].

3. **Spectral gap ($\lambda_{d,G}$):** This metric evaluates the efficiency with which information can flow through various routes in a graph. It is computed as the difference between the two largest eigenvalues of the graph $(\lambda_1 - \lambda_2)$. A large $\lambda_{d,G}$ indicates a robust graph in which information can readily traverse alternative paths, suggesting minimal bottlenecks or weak links [64].

4. **Natural connectivity ($\bar{\lambda}_G$):** Natural connectivity can be interpreted as the "average eigenvalue" of the adjacency matrix $\bar{\lambda}$, and it is defined as follows:

$$\bar{\lambda}_G = \ln\left(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda_i}\right) \tag{A5}$$

It effectively measures the redundancy of pathways through the weighted count of closed walks. This metric is closely linked to the graph's overall topology and its dynamics [65]. In simpler terms, a higher natural connectivity indicates the presence of more alternative paths, enhancing the graph's robustness against disruptions.

5. **Spectral radius (ρ_G):** The largest eigenvalue, λ_1, of an adjacency matrix, A, is called the spectral radius, ρ_G. It is closely linked to the graph's capacity to manage information flow through its paths and loops [64]. In simple terms, the greater the number of distinct paths between nodes, the better connected the graph is. A graph with many loops and alternative routes will exhibit a larger spectral radius. Graphs with a high spectral radius tend to be more resilient against failures or attacks, as information can continue to flow efficiently via alternate paths.

6. **Effective conductance (C_G):** The effective resistance R_G quantifies the robustness of a graph by accounting for both the number of parallel paths and the length of each path between node pairs. Specifically, the R_{ij} between nodes i and j is defined as the potential difference between these nodes when a unit current is injected at nodes i and withdrawn at j. The effective graph resistance R_G is the sum of R_{ij} over all pairs of nodes in the graph. An effective method for computing R_G is to express it in terms of the eigenvalues:

$$R_G = n\sum_{i=1}^{n-1} \frac{1}{\mu_i} \tag{A6}$$

where μ_i is the ith non-zero eigenvalue of L. In this work, we employ a normalized version of R_G, termed conductance, which is defined as follows:

$$C_G = \frac{n-1}{R_G} \tag{A7}$$

with $0 \leq C_G \leq 1$; a larger value of C_G indicates a higher level of robustness [67].

7. **Algebraic connectivity ($\mu_{2,G}$):** The second smallest eigenvalue of the Laplacian matrix, also known as algebraic connectivity $\mu_{2,G}$ or the Fiedler value, is a crucial measure of graph robustness [66]. Because the L is symmetric and positive semidefinite, with

each row summing to zero, its eigenvalues are real and non-negative. The smallest eigenvalue is always zero, and its multiplicity corresponds to the number of connected components in the graph. A higher $\mu_{2,G}$ indicates a more robust graph, meaning that networks with greater algebraic connectivity are more resistant to disconnection.

Table A1. Prompting strategies (A–C) used to probe how instruction formulation affects output quality.

	Prompts
A	Objective: write a fully original abstract (maximum 350 words) for a hypothetical research proposal.\Basis for the Proposal: the conceptualization of this research proposal must emerge from the critical application and interpretation of the "Focus Query" applied to the "Reference Document/Context" provided.\Key Emphasis: The abstract should build a compelling narrative that explicitly highlights an underexplored angle, an unexamined relationship, or a distinct perspective in the field, grounded in the analysis of the provided inputs.\Abstract Format:\It must be a self-contained summary, understandable on its own.\It should be written as a single continuous paragraph, with no subheadings or separate sections.\It must contain no more than 350 words (integrated within the single paragraph):\When writing the abstract, ensure the following elements are fluently included, derived from the analysis of the Document/Query:\Context and Problem (Background): briefly position the research area and specify the problem or central question the proposal aims to address, establishing its relevance within the broader field.\Approach and Methods (Methods): concisely describe the overall approach and/or the main methodological or analytical strategies planned to address the problem.\Expected results: mention the main findings that are reasonably anticipated based on the proposed approach.\Original contribution (conclusion/contribution): state the expected interpretation of the results and, crucially, how this proposal provides a new, complementary, underexplored, or distinct perspective compared to existing work, emphasizing the originality derived from the analysis of the document/query.\Required inputs (variables):\the abstract will be generated using the following information:\Reference Document/Context: context\Focus Query: query\Strict Output Constraint:\nYou must only return the text of the generated abstract, adhering to all formatting, content, word limit, and emphasis requirements. Do not include headings, titles, explanations, preambles, bullet points, markdown formatting, or any additional text. The output must consist solely of the abstract in one paragraph of plain text. Avoid "All rights reserved".
B	Before drafting the abstract, carefully analyze the 'Reference Document/Context' and 'Focus Query' to identify underexplored gaps or novel relationships. Summarize the core problem or question from the Document/Query, noting its significance and current limitations in the field. Pinpoint at least one distinct angle, unresolved tension, or overlooked methodological approach suggested by the document/query. Outline methods tailored to address this gap, ensuring that they align with the proposed originality. Project how results might challenge, refine, or expand existing knowledge.\Synthesize these steps into a single, fluid abstract (no more than 350 words) that\opens with the problem's context and relevance, anchored to the Document/Query.\Proposes methods explicitly designed to explore the identified gap.\Forecasts results and\emphasizes their novelty (e.g., 'This study is the first to...' or 'By reframing X as Y, we expect to...').\Avoids filler phrases, headings, or meta-commentary. Output only the plain-text abstract. Avoid 'All rights reserved'.\Reference Document/Context: context\Focus Query: query\n
C	You are a postdoctoral researcher in metallurgy with expertise in advanced materials characterization and processing. Your task is to write a fully original abstract (maximum 350 words) for a hypothetical research proposal derived exclusively from the provided reference document/context and focus query. The abstract must be a self-contained, single-paragraph summary with no external knowledge, subheadings, or formatting. It should include (1) context and problem: briefly situate the research question within the field, highlighting its relevance; (2) approach and methods: describe the proposed methodology, derived solely from the context; (3) expected results: anticipate key findings based on the context; and (4) original contribution: explicitly state how the proposal offers a novel or underexplored perspective, justified only by the provided inputs. Do not infer, assume, or extrapolate beyond the given material. Output only the abstract text, with no preamble, headings, or extraneous text. Avoid 'All rights reserved'.\Reference document/context: context\Focus query: query

Mean of differences ($\bar{D}$)

Considering D_j as the difference computed at iteration j, where $j \in \{1, 2, \ldots, 31\}$, the mean difference $\bar{D}$ is calculated as follows:

$$\bar{D} = \frac{1}{31} \sum_{j=1}^{31} D_j$$

Standard deviation of differences (σ_D)

The standard deviation of the differences σ_D is given as follows:

$$\sigma_D = \sqrt{\frac{1}{30} \sum_{j=1}^{31} (D_j - \bar{D})^2}$$

These formulas quantify and analyze metric variations, providing a solid basis for comparative evaluation of experimental configurations.

Experimental Server

All experiments configurations (See Table A3) were executed on a single node with the hardware and software configuration summarized in Table A2. This machine hosts both the vector database and the embedding and LLMs used in the pipeline. GPU acceleration is reserved for LLM inference, while retrieval and preprocessing stages run on the CPU.

Figure A7 displays the normalized distribution of metrics for the retrieved abstracts under the optimal configuration. This visualization is key to understanding the overall behavior of the metrics across the 31 iterations.

In contrast, Figure A8 shows a radar chart of the best individual result obtained among the 31 iterations of the optimal configuration. A substantial improvement is observed in most metrics for the generated abstract compared to the average of the retrieved abstracts. Notably, the average edge betweenness metric did not exhibit improvement in this experiment, a consistent pattern observed throughout the evaluations.

Table A2. Hardware and software specifications of the server used for all experiments.

Component	Specification
Operating system	Ubuntu 22.04.5 LTS (Jammy Jellyfish)
Kernel	Linux `6.8.0-65-generic`
CPU	Intel Core i7-14700K (20 cores, 28 threads; base 3.4 GHz, *turbo* up to 5.6 GHz; 33 MB Smart Cache)
RAM	128 GiB DDR5 3200 MHz (125 GiB visible to OS)
GPU	NVIDIA GeForce RTX 4080 (AD103, 16 GiB GDDR6X)
Storage	NVMe PCIe 4.0 ×4 SSD (1 TB)
Python environment	CPython 3.12.6 + `pip` 24.2
Key libraries	`PyTorch 2.5.1`, `Transformers 4.52.4`, `faiss-cpu 1.8`, `networkx 3.4.2`, `nvidia-cuda-toolkit 12.4.127`

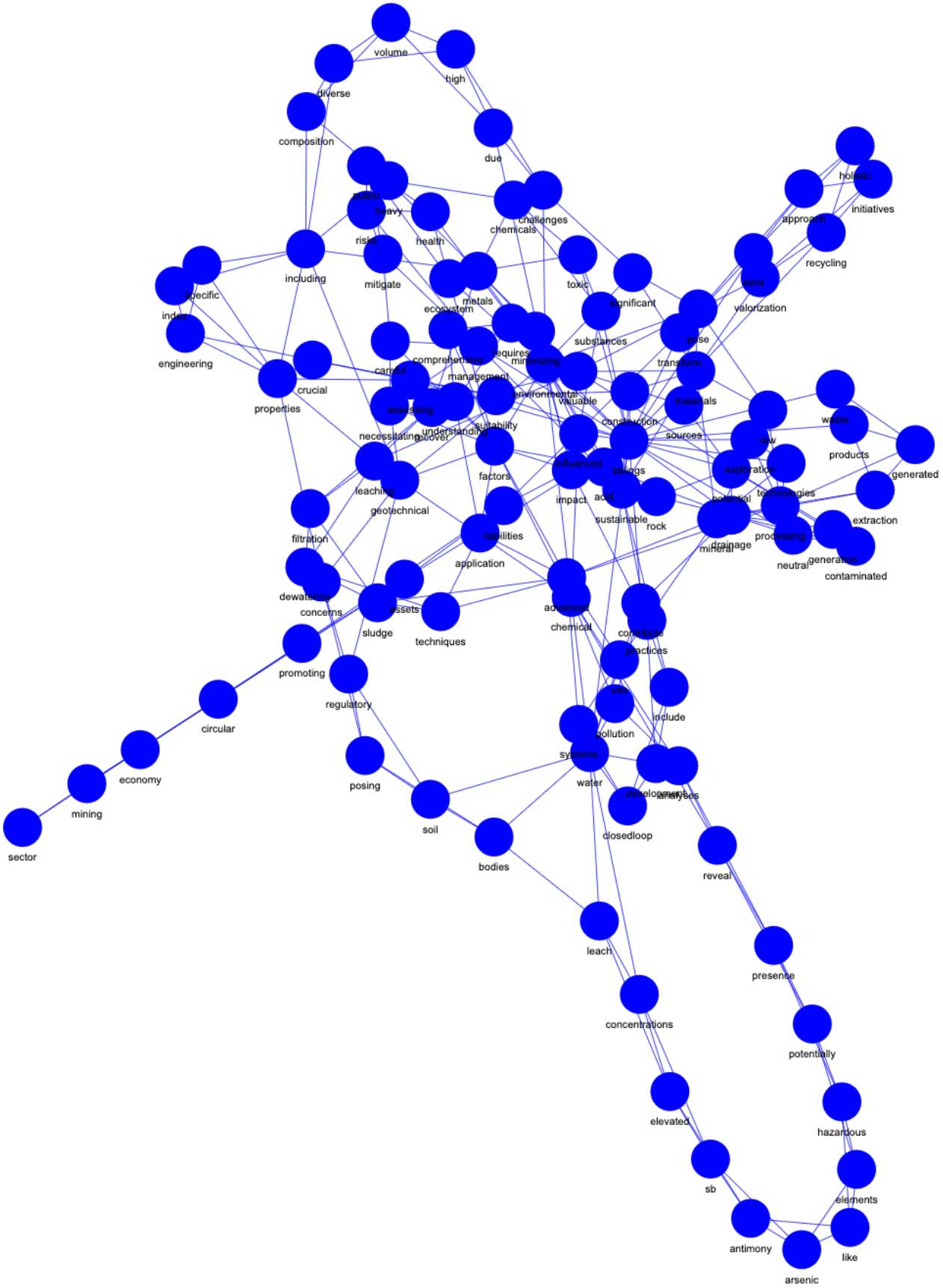


Figure A3. The figure shows a representative example of the graph structure used in Experiment 1.

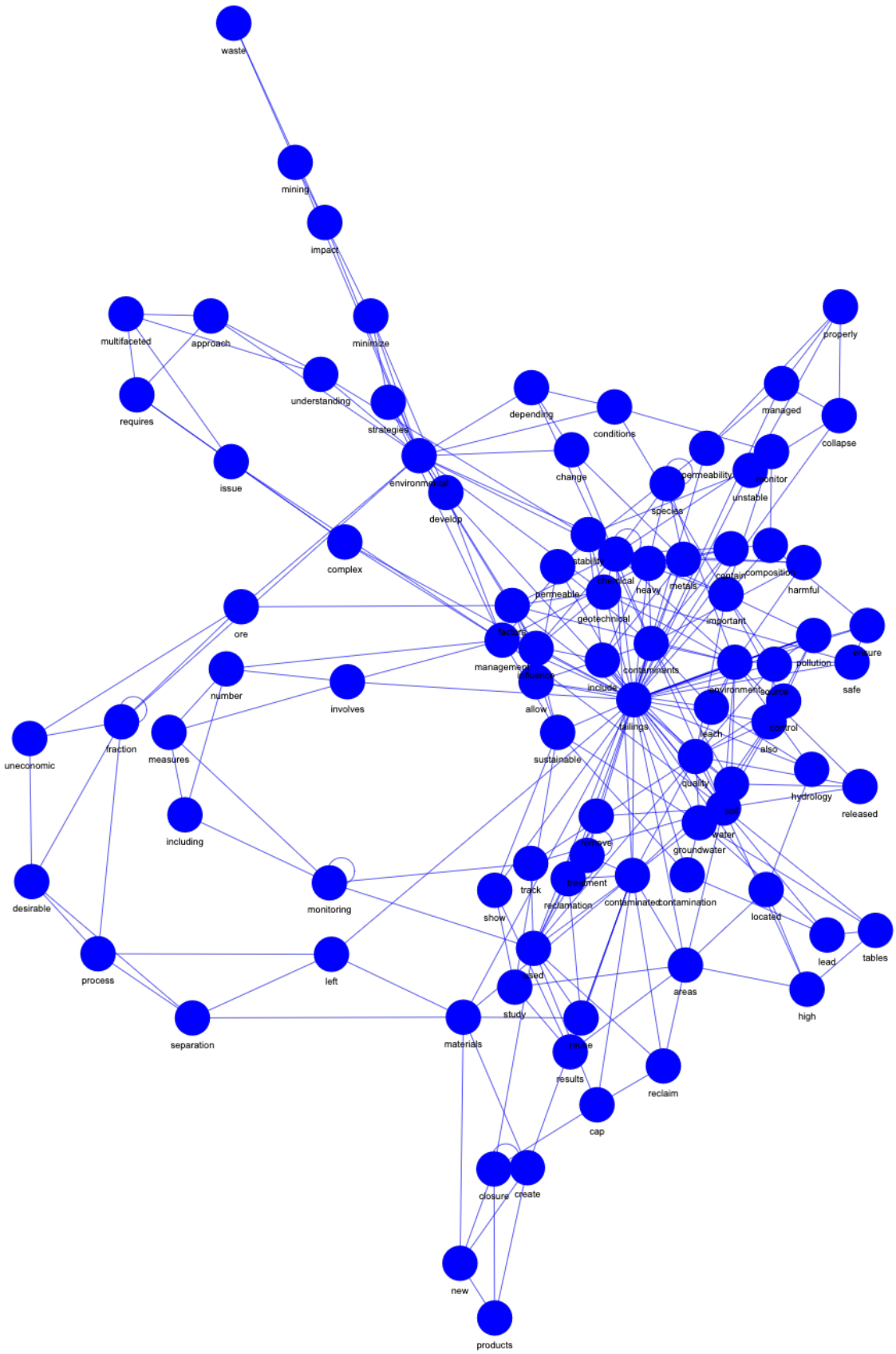


Figure A4. The figure illustrates an example of a graph generated with the optimal configuration obtained in Experiment 2.

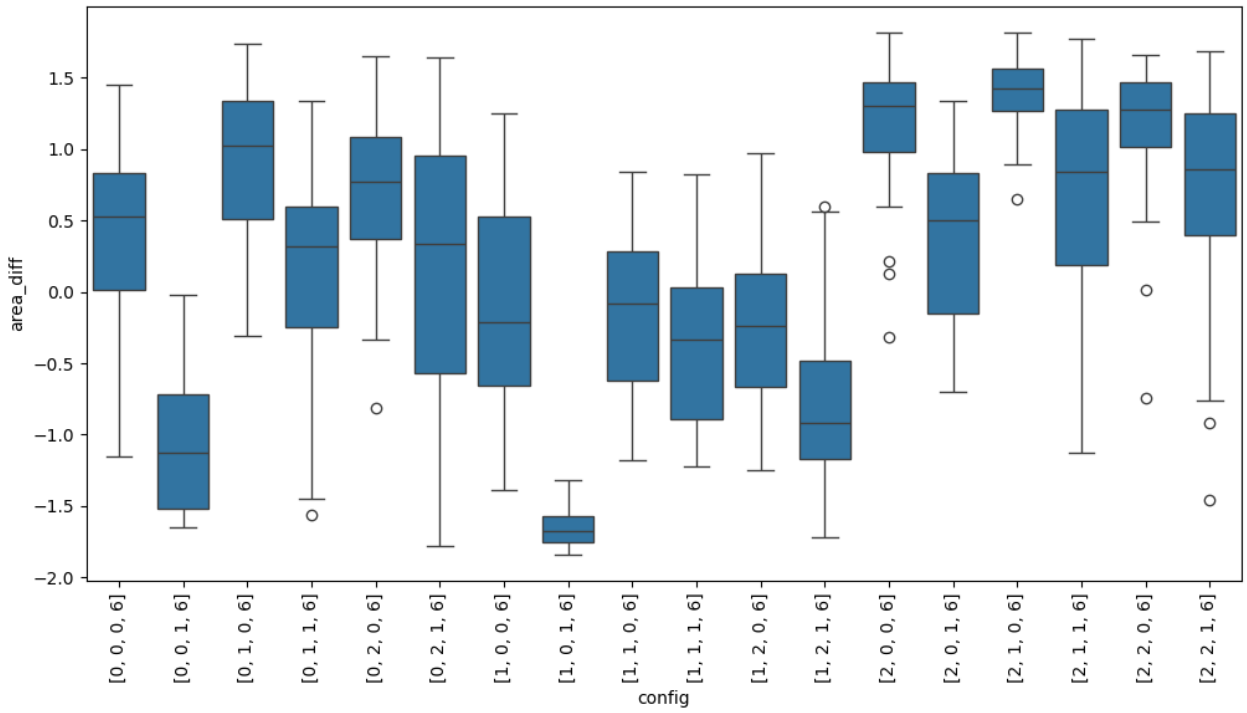


Figure A5. Distribution of the area differences between the new abstract and the 10 retrieved abstracts across 31 iterations for each configuration, with a fixed temperature of 0.7. The X-axis represents the experimental configurations defined by $[w, x, y, z]$ (where w represents the LLM model, x the embedding model, y the prompt, and z the temperature). The Y-axis shows the area difference value, calculated as area_new − area_avg, where area_new is the metric chart area of the generated abstract and area_avg is the average area of the N retrieved abstracts. The small circles represent outliers in the boxplots, corresponding to data points that fall outside the interquartile range.

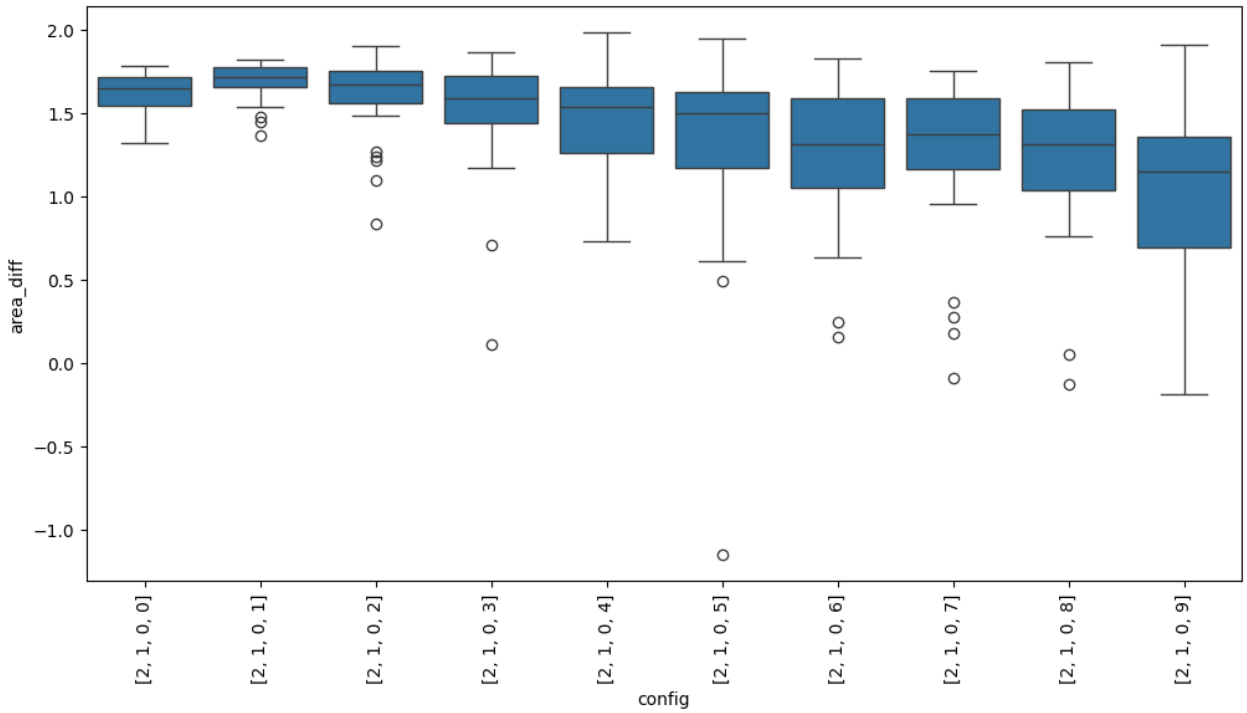


Figure A6. Distribution of radar metric area differences across 31 iterations of configuration $[2, 1, 0, z]$ evaluated at varying temperature settings. The small circles represent outliers in the boxplots, corresponding to data points that fall outside the interquartile range.

Table A3. Experimental configuration: each component defines a distinct dimension of model parametrization (see Table A1 for prompts). Configurations were represented as $[w, x, y, z]$ on all experiments.

Component	Available options

w—LLM	`0-meta-llama/Llama-3.2-3B-Instruct`
	`1-Qwen/Qwen2.5-3B-Instruct`
	`2-google/gemma-2b-it`

x—Embedding model	`0-sentence-transformers/all-mpnet-base-v2` `1-dunzhang/stella_en_400M_v5`
y—Prompt type	0—Prompt A 1—Prompt B 2—Prompt C
z—Temperature τ	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0

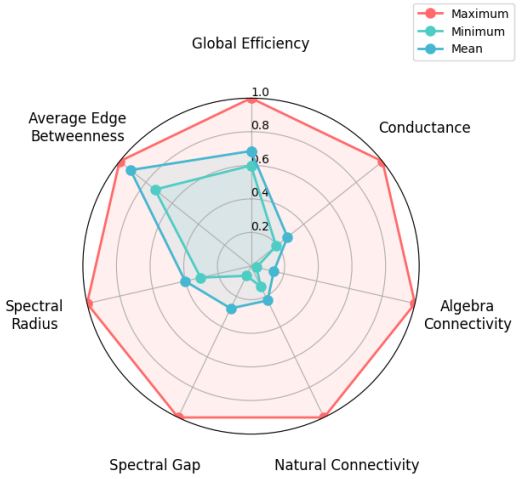


Figure A7. Normalized distribution of metrics across the 31 iterations of the best-performing configuration.

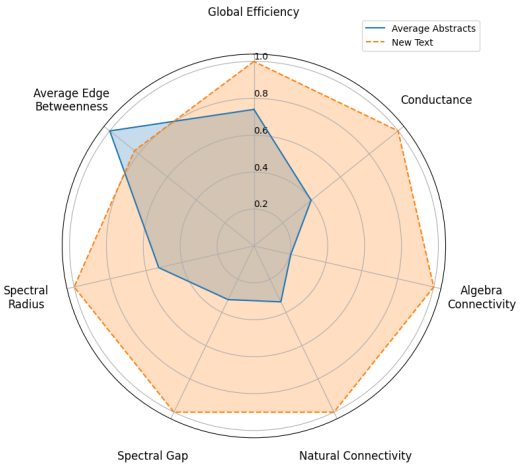


Figure A8. Comparison between the average metrics of the retrieved abstracts and those of the proposed abstract in the best-performing experiment.

References

1. Biemann, C.; Roos, S.; Weihe, K. Quantifying Semantics using Complex Network Analysis. In Proceedings of the COLING 2012, Mumbai, India, 8–15 December 2012; pp. 263–278.
2. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2025**, arXiv:2404.16130.

3. Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; Halappanavar, M.; Rossi, R.A.; Mukherjee, S.; Tang, X.; et al. Retrieval-Augmented Generation with Graphs (GraphRAG). *arXiv* **2025**, arXiv:2501.00309.
4. Havemann, F.; Scharnhorst, A. Bibliometric Networks. *arXiv* **2012**, arXiv:1212.5211. [CrossRef]
5. Patel, A.; Summers, J.; Kumar, P.; Edwards, S. Investigating the Use of Concept Maps and Graph-Based Analysis to Evaluate Learning. In Proceedings of the 2024 ASEE Annual Conference & Exposition, Portland, OR, USA, 23–26 June 2024; Available online: https://www.scopus.com (accessed on 1 July 2025).
6. Cohan, A.; Goharian, N. Revisiting Summarization Evaluation for Scientific Articles. *arXiv* **2016**, arXiv:1604.00400. [CrossRef]
7. Lapata, M.; Barzilay, R. Automatic Evaluation of Text Coherence: Models and Representations. In Proceedings of the IJCAI 2005, Edinburgh, UK, 30 July–5 August 2005; pp. 1085–1090.
8. Zhao, W.; Strube, M.; Eger, S. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 3865–3883. [CrossRef]
9. Pogorilyy, S.; Kramov, A. Assessment of Text Coherence by Constructing the Graph of Semantic, Lexical, and Grammatical Consistancy of Phrases of Sentences. *Cybern. Syst. Anal.* **2020**, *56*, 893–899. [CrossRef]
10. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *arXiv* **2024**, arXiv:2405.06211. [CrossRef]
11. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2021**, arXiv:2005.11401. [CrossRef]
12. Ayala, O.; Bechard, P. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 16–21 June 2024; Association for Computational Linguistics: Stroudsburg, PA, USA , 2024; Volume 6: Industry Track, pp. 228–238. [CrossRef]
13. Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; Nanayakkara, S. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1–17. [CrossRef]
14. Kang, M.; Kwak, J.M.; Baek, J.; Hwang, S.J. Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation. *arXiv* **2023**, arXiv:2305.18846.
15. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv* **2024**, arXiv:2402.13178. [CrossRef]
16. Luo, L.; Zhao, Z.; Haffari, G.; Phung, D.; Gong, C.; Pan, S. GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation. *arXiv* **2025**, arXiv:2502.01113. [CrossRef]
17. Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; Tang, S. Graph Retrieval-Augmented Generation: A Survey. *arXiv* **2024**, arXiv.2408.08921. [CrossRef]
18. Wang, X.; Chen, G. Complex networks: Small-world, scale-free and beyond. *IEEE Circuits Syst. Mag.* **2003**, *3*, 6–20. [CrossRef]
19. Borge-Holthoefer, J.; Arenas, A. Semantic Networks: Structure and Dynamics. *Entropy* **2010**, *12*, 1264–1302. [CrossRef]
20. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3606–3611. [CrossRef]
21. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [CrossRef]
22. Yang, R.; Yang, B.; Feng, A.; Ouyang, S.; Blum, M.; She, T.; Jiang, Y.; Lecue, F.; Lu, J.; Li, I. Graphusion: A RAG Framework for Knowledge Graph Construction with a Global Perspective. *arXiv* **2025**, arXiv:2410.17600.
23. Lehmann, F. Semantic networks. *Comput. Math. Appl.* **1992**, *23*, 1–50. [CrossRef]
24. Ma, N.; Politowicz, A.; Mazumder, S.; Chen, J.; Liu, B.; Robertson, E.; Grigsby, S. Semantic Novelty Detection in Natural Language Descriptions. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 866–882. [CrossRef]
25. Jeon, D.; Lee, J.; Ahn, J.M.; Lee, C. Measuring the novelty of scientific publications: A fastText and local outlier factor approach. *J. Inf.* **2023**, *17*, 101450. [CrossRef]
26. Ferrer-i Cancho, R.; Sole, R. The small world of human language. *Proc. Biol. Sci./R. Soc.* **2001**, *268*, 2261–2265. [CrossRef]
27. Masucci, A.; Rodgers, G. Network properties of written human language. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2006**, *74*, 026102. [CrossRef] [PubMed]
28. Pereira, H.; Fadigas, I.; Senna, V.; Moret, M. Semantic networks based on titles of scientific papers. *Phys. A Stat. Mech. Its Appl.* **2011**, *390*, 1192–1197. [CrossRef]
29. Amancio, D.R.; Machicao, J.; Quispe, L.V.C. Probing the statistical properties of enriched co-occurrence networks. *arXiv* **2024**, arXiv:2412.02664. [CrossRef]

30. Serrano, M.; Boguñá, M.; Vespignani, A. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 6483–6488. [CrossRef]
31. Quispe, L.; Tohalino, J.; Amancio, D. Using virtual edges to improve the discriminability of co-occurrence text networks. *Phys. A Stat. Mech. Its Appl.* **2021**, *562*, 125344. [CrossRef]
32. Wang, K.; Ding, Y.; Han, S. Graph neural networks for text classification: A survey. *Artif. Intell. Rev.* **2024**, *57*, 190. [CrossRef]
33. Bullinaria, J.; Levy, J. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. Res. Methods* **2007**, *39*, 510–526. [CrossRef]
34. Levy, O.; Goldberg, Y. Dependency-Based Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 2, pp. 302–308. [CrossRef]
35. Church, K.; Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **2002**, *16*, 76–83. [CrossRef]
36. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.* **1993**, *19*, 61–74.
37. Scopus. 2025. Available online: https://www.scopus.com (accessed on 1 July 2025).
38. Wang, X.; Koç, Y.; Derrible, S.; Ahmad, S.N.; Pino, W.J.; Kooij, R.E. Multi-criteria robustness analysis of metro networks. *Phys. A Stat. Mech. Its Appl.* **2017**, *474*, 19–31. [CrossRef]
39. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997. [CrossRef]
40. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084. [CrossRef]
41. Kusupati, A.; Bhatt, G.; Rege, A.; Wallingford, M.; Sinha, A.; Ramanujan, V.; Howard-Snyder, W.; Chen, K.; Kakade, S.; Jain, P.; et al. Matryoshka Representation Learning. *arXiv* **2024**, arXiv:2205.13147.
42. Zhang, D.; Li, J.; Zeng, Z.; Wang, F. Jasper and Stella: Distillation of SOTA embedding models. *arXiv* **2025**, arXiv:2412.19048.
43. Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive Text Embedding Benchmark. *arXiv* **2023**, arXiv:2210.07316. [CrossRef]
44. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv* **2023**, arXiv:2308.03281. [CrossRef]
45. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
46. Team, Q. *Qwen2.5: A Party of Foundation Models*; GitHub: San Francisco, CA, USA , 2024.
47. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen Technical Report. *arXiv* **2023**, arXiv:2309.16609. [CrossRef]
48. Team, G. *Gemma*; Kaggle: San Francisco, CA, USA, 2024. [CrossRef]
49. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv* **2022**, arXiv:2110.08207. [CrossRef]
50. Weller, O.; Durme, B.V.; Lawrie, D.; Paranjape, A.; Zhang, Y.; Hessel, J. Promptriever: Instruction-Trained Retrievers Can Be Prompted Like Language Models. *arXiv* **2024**, arXiv:2409.11136. [CrossRef]
51. Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q.V.; Zhou, D.; Chen, X. Large Language Models as Optimizers. *arXiv* **2024**, arXiv:2309.03409. [PubMed]
52. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
53. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
54. Wang, Z.M.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Yang, J.; et al. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. *arXiv* **2024**, arXiv:2310.00746.
55. Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; Dong, X. Better Zero-Shot Reasoning with Role-Play Prompting. *arXiv* **2024**, arXiv:2308.07702.
56. Vijayarani, S.; Janani, R. Text mining: Open source tokenization tools-an analysis. *Adv. Comput. Intell. Int. J. (ACII)* **2016**, *3*, 37–47.
57. Sun, Y.; Platoš, J. A method for constructing word sense embeddings based on word sense induction. *Sci. Rep.* **2023**, *13*, 12945. [CrossRef] [PubMed]
58. Ma, R.; Jin, L.; Liu, Q.; Chen, L.; Yu, K. Addressing the polysemy problem in language modeling with attentional multi-Sense embeddings. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020. [CrossRef]
59. Saeidi, M.; Milios, E.; Zeh, N. Biomedical Word Sense Disambiguation with Contextualized Representation Learning. In Proceedings of the WWW '22: Companion Proceedings of the Web Conference 2022, Virtual Event, 25–29 April 2022; pp. 843–848. [CrossRef]

60. Hagberg, A.; Conway, D. Networkx: Network Analysis with Python. 2020, pp. 1–48. Available online: https://networkx.github.io (accessed on 1 July 2025).
61. Liu, H.; Cong, J. Language clustering with word co-occurrence networks based on parallel texts. *Chin. Sci. Bull.* **2013**, *58*, 1139–1144. [CrossRef]
62. Latora, V.; Marchiori, M. Efficient behavior of small-world networks. *Phys. Rev. Lett.* **2001**, *87*, 198701. [CrossRef]
63. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, *40*, 35–41. [CrossRef]
64. Van Mieghem, P. *Graph Spectra for Complex Networks*; Cambridge University Press: Cambridge, UK, 2023.
65. Jun, W.; Barahona, M.; Yue-Jin, T.; Hong-Zhong, D. Natural connectivity of complex networks. *Chin. Phys. Lett.* **2010**, *27*, 078902. [CrossRef]
66. Fiedler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, *23*, 298–305. [CrossRef]
67. Ellens, W.; Spieksma, F.M.; Van Mieghem, P.; Jamakovic, A.; Kooij, R.E. Effective graph resistance. *Linear Algebra Its Appl.* **2011**, *435*, 2491–2506.
68. Warrens, M.J. Cohen's linearly weighted kappa is a weighted average. *Adv. Data Anal. Classif.* **2012**, *6*, 67–79. [CrossRef]
69. Renze, M. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 7346–7356.
70. Star, S.L.; Griesemer, J.R. Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Soc. Stud. Sci.* **1989**, *19*, 387–420. [CrossRef]
71. Barbierato, E.; Gatti, A.; Incremona, A.; Pozzi, A.; Toti, D. Breaking Away From AI: The Ontological and Ethical Evolution of Machine Learning. *IEEE Access* **2025**, *13*, 55627–55647. [CrossRef]
72. Basile, P.; Siciliani, L.; Musacchio, E.; Semeraro, G. Exploring the Word Sense Disambiguation Capabilities of Large Language Models. *arXiv* **2025**, arXiv:2503.08662. [CrossRef]
73. Toshevska, M.; Stojanovska, F.; Kalajdjieski, J. Comparative Analysis of Word Embeddings for Capturing Word Similarities. In Proceedings of the 6th International Conference on Natural Language Processing (NATP 2020), Copenhagen, Denmark, 25–26 April 2020; Aircc Publishing Corporation: Chennai, India, 2020; NATP 2020. pp. 9–24. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.