ResearchFusion AI: An Evidence-Grounded Framework for Multi-Paper Research Gap Analysis, Novelty Assessment, and Roadmap Generation

Rohan Gaikwad[1], Ishita Deepak Shete[2], Manish Narayan Gholap[3], Shrikar Amol Dehadray[4]
Aarti Pimpalkar[5], Mayuresh Gulame[6]
[1,2,3,4,5,6]*Department of CSE, School of Computing, MIT Art, Design & Technology University, Pune, India*

***Abstract*—Scientific research requires researchers to compare prior studies before defining a meaningful research problem. This process involves examining methods, datasets, evaluation measures, results, limitations, and future-work statements across multiple papers. This paper proposes ResearchFusion AI, a multi-domain framework designed to support that process for a user-selected collection of three to ten research papers. The framework organizes extracted information into structured records, generates cross-paper comparisons, groups related limitations, and identifies research dimensions that appear underexplored in the supplied corpus. It then produces candidate research directions and a heuristic novelty indicator on a 0–100 scale, based on semantic distance, uncommon combinations of techniques, coverage of selected gaps, and a technical-contribution rubric. A separate feasibility estimate communicates practical constraints. Evidence-linked retrieval and schema-constrained outputs are used to make generated statements traceable to source passages. The paper describes the proposed architecture, six processing modules, a five-phase research roadmap, and an evaluation protocol. A healthcare example illustrates the workflow; it is not an empirical performance result. ResearchFusion AI is intended as a decision-support system, and its generated ideas and novelty estimates require expert review and a broader literature search.**

***Index Terms*—evidence-grounded generation, large language models, literature review, novelty assessment, research gap analysis, research planning, retrieval-augmented generation**

I. INTRODUCTION

The growth of scientific publications makes it increasingly difficult to build a reliable overview of a research area. Before proposing a new method, researchers typically identify relevant studies, extract their problem definitions and experimental settings, compare reported findings, and examine limitations. When this work is performed manually, important details may be overlooked and comparisons can become inconsistent, particularly when papers use different datasets or evaluation protocols.

Language models and scientific-text representations offer tools for assisting with document analysis [1], [2], [5]–[8]. However, summarizing an individual paper does not by itself establish how several studies relate to one another, which limitations recur, or whether a proposed direction addresses a defensible gap. Generated statements can also be plausible but unsupported, so a research-support workflow must preserve links between its conclusions and the source material [9], [10].

This paper proposes ResearchFusion AI, a modular framework for analyzing a user-selected set of three to ten research papers across domains such as healthcare, defence, environment, education, and business. Its workflow covers paper ingestion, structured extraction, comparative analysis, research-gap identification, candidate-idea generation, heuristic novelty assessment, and roadmap preparation. The framework is designed to assist researchers rather than replace expert judgment.

- A modular workflow connects paper extraction, comparison, gap analysis, idea generation, novelty assessment, and roadmap planning;

- A traceable gap-identification process groups stated limitations and distinguishes them from inferred observations;
- A decomposable novelty indicator combines four defined components, while feasibility is reported separately;
- A retrieval-grounded generation process attaches source evidence to extracted information and candidate research directions;
- A proposed architecture and evaluation protocol define how the framework can be implemented and assessed.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the problem formulation, architecture, and six modules. Section IV provides an illustrative use case. Section V specifies an evaluation protocol. Section VI discusses limitations and responsible use, and Section VII concludes the paper.

II. RELATED WORK

A. Scientific Document Processing

Scientific document-processing resources support the conversion of research papers into machine-readable text. The S2ORC corpus provides a large collection of structured scholarly documents [3], while GROBID extracts bibliographic metadata and document structure from scholarly PDFs [4]. Scientific language representations such as SciBERT and SPECTER support domain-aware text representation [5], [6], and Sentence-BERT provides sentence embeddings suitable for semantic comparison [7]. ResearchFusion AI uses these types of tools as enabling components; it does not claim to introduce a new document encoder.

B. Language Models and Retrieval-Augmented Generation

Pretrained language models such as BERT and GPT-3 demonstrate the ability of large-scale models to perform varied language tasks [2], [8]. Retrieval-augmented generation conditions generated text on retrieved source passages [9], which can improve traceability when the retrieval context is relevant. Nevertheless, language models may generate unsupported or incorrect content [10]. For this reason, the proposed workflow requires evidence spans for extracted fields and reported gaps, validates structured outputs, and marks unavailable information rather than silently filling it.

C. Idea Generation and Novelty Assessment

Recent work has explored language models as assistants for scientific ideation. ResearchAgent combines literature retrieval with iterative review agents to propose and refine research ideas [11]. The AI Scientist explores a more automated research workflow that includes idea generation, coding, experimentation, and manuscript preparation [12]. A large-scale study by Si et al. reports that LLM-generated ideas were rated as more novel but somewhat less feasible than ideas from expert researchers, highlighting the need to assess feasibility separately from novelty [13]. ResearchFusion AI focuses on a narrower decision-support task: comparing a user-supplied paper set, linking candidate gaps to evidence, and translating a selected direction into a staged plan.

D. Positioning

Table I provides a qualitative comparison of common literature-analysis approaches. Unlike a single-paper summary, ResearchFusion AI is designed to connect multiple papers to evidence-linked gap candidates, candidate ideas, and an implementation roadmap. The comparison describes intended capabilities and is not the result of a controlled product benchmark.

TABLE I. COMPARISON OF LITERATURE-ANALYSIS APPROACHES

Capability	Single-paper summarizers	Manual review	ResearchFusion AI
Compares many papers	Limited	Yes	Yes (3–10 papers)
Finds research gaps	No	Depends on reader	Evidence-linked candidates
Suggests new ideas	No	Depends on reader	Grounded suggestions
Novelty score (0–100)	No	No	Heuristic (0–100)
Feasibility check	No	Depends on reader	Separate estimate

Capability	Single-paper summarizers	Manual review	ResearchFusion AI
Roadmap generation	No	Depends on reader	Phase-wise plan
Typical time[a]	Minutes per paper	Days to weeks	Not measured

[a]Time requirements vary by task and corpus; no controlled timing results are reported for Research Fusion AI.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. Problem Formulation

Let $P = \{p_1, \ldots, p_n\}$, where $3 \leq n \leq 10$, denote the uploaded papers; let d denote the selected domain and q an optional topic. For each paper p_i, the framework aims to produce a structured record r_i containing bibliographic details, problem statement, method, dataset, metrics, results, and limitations, with supporting evidence where available. Across P, it then constructs (i) a normalized comparison, (ii) a ranked set of evidence-linked gap candidates G, (iii) candidate ideas X, and (iv) a roadmap $\Phi(x^*)$ for a user-selected idea x^*. Each candidate idea receives a heuristic novelty indicator N(x) in [0, 100] and a separate feasibility estimate F(x) in [0, 1]. The objective is to support faster, more transparent analysis while keeping researchers responsible for verification and final decisions.

B. Architecture Overview

Fig. 1 presents the proposed architecture. A user interface accepts PDF files, a research topic, and a domain selection. A backend service coordinates document parsing, extraction, comparison, retrieval, and report generation. An LLM and an embedding model are accessed through replaceable provider interfaces, while a shared evidence store maintains source text, page references, structured records, and intermediate outputs. The six modules exchange validated, typed data objects so that individual components can be updated independently. Domain-specific checklists and metric vocabularies are supplied through configuration rather than hard-coded into the processing logic.

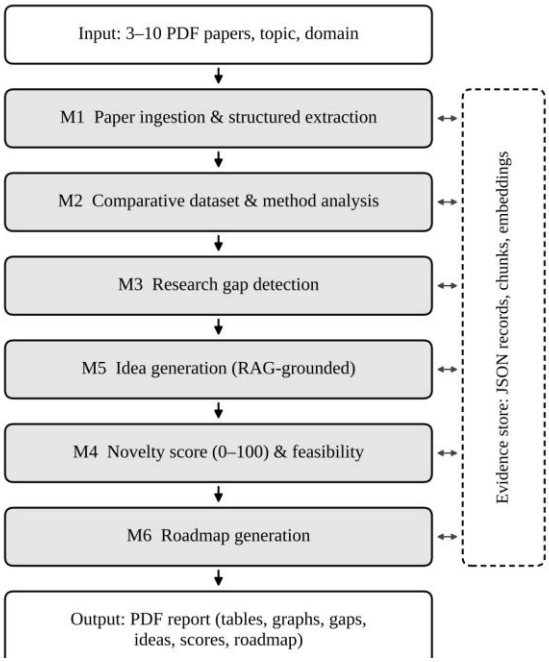


Fig. 1. Proposed ResearchFusion AI architecture. The six modules exchange structured outputs and source references through a shared evidence store.

C. Module 1: Paper Ingestion and Structured Extraction

Each uploaded PDF is converted into text while preserving available section boundaries, page references, tables, and bibliographic information. A structured parser such as GROBID may be used where suitable, with a fallback text extractor for files that cannot be processed by the primary parser [4]. The extracted text is divided into section-aware chunks for retrieval. A language model populates a fixed schema containing metadata, problem, method, dataset, results, metrics, limitations, and future work. Each populated field should include its supporting passage and page or section reference, together with a confidence or validation status. Schema validation is applied before the record is passed to later modules; missing information is recorded as unavailable rather than guessed.

TABLE II. STRUCTURED EXTRACTION SCHEMA

Group	Fields	Used by
Metadata	Title, authors, year, keywords, source	Report, M2
Content	Abstract, research problem, method, evidence span	M2, M5

Group	Fields	Used by
Data	Dataset name, size, type, split and licence (if stated)	M2, M6
Results	Metric names, protocol and reported values	M2
Limitations	Author-stated limitations and future work, with source passage	M3

D. Module 2: Comparative Dataset and Method Analysis

The comparison module normalizes names of methods, datasets, and evaluation metrics so that equivalent terms can be aligned across records. It presents side-by-side summaries of methods, dataset characteristics, and reported measures such as Accuracy, Precision, Recall, and F1-score. Numerical results are compared only when the dataset, split, and evaluation protocol are sufficiently compatible. Where these conditions differ or are not reported, the system labels the values as non-comparable instead of producing a misleading ranking. Missing values remain explicitly marked.

E. Module 3: Research Gap Detection

The gap-detection module combines limitations stated by authors, recurring issues across papers, and research dimensions not observed in the supplied set. It performs the following steps while preserving the distinction between direct evidence and system-generated inference.

1) Limitation collection: Stated limitations and future-work statements are extracted from each paper and stored with their source passages. If a paper does not state a limitation, the system may generate a separate inferred observation from its method or evaluation design. Such observations are explicitly labeled as inferred and are not presented as quotations or direct author claims.
2) Clustering: Limitation statements are embedded using a sentence-embedding model such as Sentence-BERT [7]. Agglomerative clustering with cosine distance groups semantically similar statements. For each cluster c, support s(c) is the number of distinct papers contributing a statement. The support count is reported as a property of the uploaded set, not as proof that the issue is prevalent across the entire research field.
3) Coverage check: A domain-specific checklist may include explainability, external validation, dataset diversity, computational cost, privacy, and reproducibility. The system checks whether these dimensions are discussed in the uploaded papers and reports dimensions that are not observed. Absence from the supplied papers is described as an evidence gap in the corpus, not as proof that the wider field has ignored the topic.
4) Gap ranking: Candidate gaps from the preceding steps are consolidated and ranked using cross-paper support, relevance to the selected topic, and evidence quality. Each reported gap includes the contributing papers and supporting passages. Directly stated limitations and inferred observations remain visibly distinguished in the report.

F. Module 4: Novelty Score Estimation

Module 4 assigns a heuristic score to an idea x, whether generated by Module 5 or entered by a user. The idea is represented by a concise description of its problem, proposed method, and expected contribution. Let e(·) denote the embedding function and let the reference corpus contain the uploaded papers, with optional external scholarly records retrieved from a source such as Semantic Scholar [15]. The cosine similarity between the idea and a reference paper p is defined as follows:

$$\mathrm{sim}(x, p) = \frac{e(x)\cdot e(p)}{|e(x)||e(p)|} \quad (1)$$

The semantic-distance component is calculated from the k most similar reference papers:

$$D(x) = 1 - \frac{1}{k}\sum_{p\in \mathrm{Top-k}(x)} \mathrm{sim}(x, p) \quad (2)$$

Let T_x denote the set of techniques identified in idea x, and let n(a, b) count the reference papers that mention both techniques a and b. Following the view that atypical combinations can be associated with scientific impact [14], the combination component measures the proportion of technique pairs that do not co-occur in the reference corpus. If fewer than two techniques are identified, this component is set to zero and flagged as unavailable for interpretation:

$$C(x) = \frac{|\{(a,b)\in \mathrm{pairs}(T_x) : n(a,b)=0\}|}{|\mathrm{pairs}(T_x)|} \quad (3)$$

Gap coverage is the proportion of the selected research gaps G_{sel} addressed by idea x, where G_x is the subset of selected gaps covered by x. At least one gap must be selected for this component to be calculated:

$$G_{cov}(x) = \frac{|G_x|}{|G_{sel}|} \quad (4)$$

The technical-contribution component T(x) is the mean of three rubric criteria: (i) whether the idea adds a distinct component or mechanism, (ii) whether that contribution is justified by an identified gap, and (iii) whether the design can be tested and reproduced. An LLM may provide a preliminary rubric-based rating, but the rationale should be retained and the rating should be treated as provisional until reviewed by a domain expert. The four components are normalized to [0, 1], and the default weights sum to one:

$$N(x) = 100[w_D D(x) + w_C C(x) + w_G G_{cov}(x) + w_T T(x)] \quad (5)$$

TABLE III. COMPONENTS OF THE NOVELTY SCORE

Component	What it measures	Default weight
D(x)	Distance from the nearest papers in the selected reference corpus	0.35
C(x)	Share of technique pairs not observed together in that corpus	0.25
$G_{cov}(x)$	Coverage of the selected, evidence-linked research gaps	0.20
T(x)	Rubric-based assessment of technical contribution	0.20
F(x)	Feasibility estimate (reported separately from N)	–

The feasibility estimate F(x) is assessed separately using factors such as data availability, computational requirements, maturity of the required tools, and expected effort. It is not included in N(x), allowing users to distinguish a high novelty signal from a practical implementation risk. The default weights are initial design choices and must be calibrated using expert annotations. Score bands should be interpreted only as relative indicators within the selected reference corpus: a higher score does not establish global originality or prove that no similar work exists. The report should therefore display the nearest reference papers and encourage a wider prior-art search.

G. Module 5: New Idea Generation

For each high-priority gap, the idea-generation module retrieves its linked evidence passages and, when enabled, relevant external abstracts or publications [9]. The generation prompt includes the gap statement, its evidence, domain constraints, and a required output schema. Each candidate idea contains a title, target gap, proposed approach, main technical components, suggested data and evaluation procedure, and expected contribution. Near-duplicate candidates are filtered using semantic similarity. The remaining candidates are passed to Module 4 and displayed with their novelty indicator, feasibility estimate, and supporting references.

H. Module 6: Research Roadmap Generation

For a user-selected idea x^*, the roadmap module prepares an implementation plan containing candidate dataset and their sources, relevant tools and libraries, baseline methods, evaluation metrics, and planned experiments. Dataset licences and access conditions should be recorded where they can be verified; uncertain details are marked for manual confirmation. The plan is divided into five phases, with tasks, dependencies, expected outputs, and durations that can be adjusted to the researcher's schedule.

TABLE IV. DEFAULT ROADMAP PHASES

Phase	Main tasks	Deliverables
1	Define the problem, scope, research questions, and data requirements	Problem statement and data plan
2	Select and reproduce suitable baseline methods	Baseline implementation and results
3	Develop the proposed method or system	Working prototype
4	Run planned experiments, ablations, and validation	Evaluation results and analysis
5	Interpret findings, document limitations, and prepare dissemination	Final report or research paper

I. Implementation Considerations

The architecture is technology-independent and can be implemented using a web interface, an API-based backend, a document-parsing component, an LLM provider, an embedding service, and persistent storage for files and evidence records. The specific frameworks, model providers, versions, and database products must be documented from the actual deployed prototype rather than assumed from the reference architecture. This separation allows the

language-model provider or storage layer to be changed without altering the six-module workflow.

IV. ILLUSTRATIVE CASE STUDY

Consider a hypothetical healthcare user who uploads several papers on MRI-based brain-tumour detection. This example is included only to illustrate the intended workflow; it does not report an experiment or results obtained from a deployed system.

Modules 1 and 2 would extract each paper's stated classification method, MRI dataset, experimental protocol, and reported evaluation measures. The comparison would group results by dataset and protocol, while flagging missing or non-comparable information. Module 3 would organize directly stated limitations, such as restricted external validation or limited interpretability, and identify whether these topics are absent from the uploaded set. A candidate direction might, for example, investigate visual explanations using a method such as Grad-CAM [16]; however, the system would need to verify the supporting evidence and search the wider literature before presenting the direction as novel. Any inferred observation would be labeled separately from the authors' own statements.

For illustration only, suppose a candidate idea receives $D = 0.40$, $C = 0.50$, $G_{cov} = 1.00$, and $T = 0.60$. Using the default weights, its heuristic score is $N = 100 \times (0.35 \times 0.40 + 0.25 \times 0.50 + 0.20 \times 1.00 + 0.20 \times 0.60) = 58.5$. This numerical example demonstrates the calculation only; it is not a measured score or evidence of the idea's actual novelty. The roadmap would then outline dataset verification, baseline reproduction, prototype development, external validation, and reporting.

V. EVALUATION PROTOCOL

This section defines a planned evaluation protocol; no benchmark results are claimed in this paper. The evaluation set should contain collections of three to ten papers from at least two domains, with relevant fields and gaps annotated independently by domain experts. The model version, prompts, retrieval settings, and decoding parameters should be logged to support reproducibility. The five research questions and their proposed metrics are summarized in Table V.

TABLE V. EVALUATION PROTOCOL BY MODULE

Question	Metrics	Compared with
RQ1 Extraction (M1, M2)	Field-level precision, recall, F1, and proportion of values supported by cited passages	Independent annotations by two reviewers
RQ2 Gap validity (M3)	Expert-rated gap precision; inter-rater agreement (e.g., Cohen's κ)	Gaps identified through independent expert review
RQ3 Score validity (M4)	Correlation with expert novelty ratings; component ablation; weight sensitivity	Expert ratings and simpler scoring baselines
RQ4 Idea quality (M5)	Blind ratings (1–5) for novelty, feasibility, clarity, and specificity	Human-written ideas and single-prompt LLM baseline
RQ5 Usefulness (M6)	Task completion time and usefulness rating (1–5)	Manual literature review and planning

For idea-quality evaluation (RQ4), candidate ideas should be presented to reviewers without revealing whether they were generated by the system or by a comparison baseline. This blind-review approach is motivated by prior human evaluation of LLM-generated research ideas [13].

VI. DISCUSSION AND LIMITATIONS

1) Hallucination and evidence quality: Evidence spans, schema checks, and rules for marking missing fields can reduce unsupported output but cannot eliminate model errors [10]. Users should be able to inspect the source passage for every important extracted field and gap.
2) Limited input corpus: A set of three to ten papers may not represent the full research field. A gap can reflect the user's paper-selection strategy rather than a genuine absence of work. Results must therefore be described as relative to the uploaded corpus, and broader scholarly retrieval should be used before making novelty claims.
3) PDF parsing: Tables, figures, equations, and multi-column layouts may not be extracted reliably from every PDF. Users should verify

important numerical values against the original document.

4) Heuristic novelty estimate: Embedding similarity measures textual or semantic proximity, not novelty in a formal scholarly sense. The initial weights are not empirically calibrated, and the score may be sensitive to the selected papers, embedding model, and technique extraction. It should be treated as a screening aid rather than a substitute for a systematic literature or patent search.
5) Privacy and responsible use: Uploaded manuscripts may contain unpublished or confidential material. A deployed system should apply access controls, secure storage, retention policies, and clear disclosure of any external model processing. Researchers remain responsible for checking generated ideas, respecting source licences, and making final scholarly decisions.

VII. CONCLUSION AND FUTURE WORK

This paper proposed ResearchFusion AI, a modular framework for organizing and comparing multiple research papers, identifying evidence-linked gap candidates, generating candidate research directions, and preparing a phase-wise roadmap. Its novelty indicator combines semantic distance, uncommon technique combinations, selected-gap coverage, and a technical-contribution rubric, while feasibility is reported separately. The framework emphasizes traceability by linking extracted information and gap statements to source passages and by distinguishing author-stated limitations from inferred observations. The healthcare example is illustrative, and the evaluation protocol remains to be carried out. Future work includes implementing and testing the complete pipeline, building expert-annotated benchmarks, calibrating the scoring weights, measuring extraction and gap quality, and evaluating the usefulness of generated roadmaps.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[2] T. Brown et al., “Language models are few-shot learners,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[3] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, “S2ORC: The Semantic Scholar Open Research Corpus,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 4969–4983.

[4] P. Lopez, “GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in Res. Adv. Technol. Digit. Libraries (ECDL), 2009, pp. 473–474.

[5] Chinthamu, N., Pimpalkar, A. P., Vanisree, K., Sharma, P., & Kirubakaran, M. K. (2025). A Novel Approach Utilizing Artificial Intelligence and Wireless Sensor System for Ensuring Safety Through Fire Detection. Technological Applications for Smart Sensors: Intelligent Applications for Real-Time Strategies, 307.

[6] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, “SPECTER: Document-level representation learning using citation-informed transformers,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 2270–2282.

[7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. EMNLP-IJCNLP, Hong Kong, China, 2019, pp. 3982–3992.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[9] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[10] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, 2023.

[11] J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang, “ResearchAgent: Iterative research idea generation over scientific literature with large language models,” in Proc. NAACL-HLT, vol. 1, 2025, pp. 6709–6738, doi: 10.18653/v1/2025.naacl-long.342.

[12] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha, “The AI Scientist: Towards fully automated open-ended scientific discovery,” arXiv:2408.06292, 2024, doi: 10.48550/arXiv.2408.06292.

[13] C. Si, D. Yang, and T. Hashimoto, “Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2025.

[14] A. P. Pimpalkar, R. Urmude, A. Ghule, P. Jayappa, J. R. Prasad and M. Gulame, "Machine Learning-Based Stock Forecasting," 2025 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2025, pp. 1-5, doi: 10.1109/ESCI63694.2025.10988318.

[15] R. Kinney et al., “The Semantic Scholar Open Data Platform,” arXiv:2301.10140, 2023.

[16] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618–626.