International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

ResearchLens-CASCO: A Context-Adaptive Cognitive Semantic Intelligence Framework for Automated Literature Exploration Using Workflow-Orchestrated NLP Pipelines

Nirupama B, Deekshitha K R, Rakshitha M G, Sudeep Y M
4PS22CI034 | 4PS22CI007 | 4PS22CI041 | 4PS22CI054
Department of CS&E (AI & ML), P.E.S College of Engineering (PESCE), Mandya, Karnataka, India
Email: bnirupamab@gmail.com | deekshitha26kr@gmail.com | rakshithagangadhar19@gmail.com | sudeepym000@gmail.com

Prof. Renuka H R
Assistant Professor (Guide), Department of CS&E (AI & ML), P.E.S College of Engineering (PESCE), Mandya, Karnataka, India

Abstract:

There has been an exponential increase in the number of scientific publications, making the analysis of academic literature, discovery of semantic knowledge, and synthesis of research more challenging and complex. Existing research systems have been mainly dedicated to isolated tasks, such as semantic retrieval, summarization, topic modeling or conversational interaction, each executed independently, leading to scattered analytical workflows, and negligible contextual continuity within stages of research analysis.

In this paper, we propose ResearchLens-CASCO, a Context-Adaptive Semantic Cognitive Orchestration framework for intelligent automated literature analysis through workflow-orchestrated NLP-pipelines. The framework enables semantic document understanding, dynamic context propagation, adaptive workflow optimization, persistent cognitive memory, transformer-based topic modeling, contextual retrieval, and conversational research assistance through a single cognitive research intelligence architecture.

ResearchLens-CASCO has the following four major new contributions: (1) Dynamic Semantic Context Propagation (DSCP) for maintaining contextual continuity over analytical stages; (2) Adaptive Cognitive Workflow Optimization (ACWO) for on-the-fly optimization of workflow execution paths; (3) Cognitive Research State Memory (CRSM) for persistent semantic reasoning and continuity of interaction; and (4) Iterative Research Attention Engine (IRAE) for adaptive contextual literature ranking and exploration.

The backend part of the application is developed using Node.js with Express.js for API routes and the frontend uses React.js. Experimental results show an improved contextual continuity, dynamic workflow adaptivity, and upgraded literature exploration efficiency with a semantic retrieval precision of 92.4% over traditional research-analysis baselines. The framework offers a scalable path toward the next generation of AI-based cognitive research intelligence systems for automated academic knowledge discovery environments.

Keywords— Semantic NLP, Cognitive Research Intelligence, Workflow Orchestration, Adaptive Semantic Systems, BERTopic, KeyBERT, Conversational Re- trieval, Semantic Context Propagation, Literature Review Automation, Research Intelligence

I. INTRODUCTION

The rapid proliferation of digital scientific literature has reshaped the architecture of contemporary research systems. Within these, millions of papers are produced annually in areas such as Artificial Intelligence, Cybersecurity, Healthcare, Data Science, Cloud Computing, and multi-domain engineering. While this increase in access makes scientific knowledge more accessible, it also poses significant challenges in terms of semantic understanding of the literature, thematic reading, contextual retrieval, and synthesis of research.

Conventional literature-review processes are largely manual, laborious, and challenging to effectively scale over large bodies of research. Scientists have long been faced with the challenge of reading large numbers of papers in order to identify research trends, compare methods, find contradictions, assess experimental efficacy, and locate avenues for unexplored research. This cognitive load is further increased when the boundaries of the domain become fuzzy and interdisciplinary integration is needed.

Popular academic platforms such as Google Scholar, Semantic Scholar, IEEE Xplore, ResearchGate mainly offer document indexing and a keyword-based retrieval. These platforms facilitate discovery of research, but they are not

designed to incorporate cognitive semantic intelligence that maintain contextual continuity across multiple layers of analysis. Each platform is a siloed search interface, not an intelligent reasoning agent. Recent development of transformer-based NLP models, semantic embeddings, Retrieval-Augmented Generation (RAG), workflow orchestration frameworks and conversational AI systems have made the leaps in the ability of automated document understanding. A number of models that provide good semantic clusters, contextual retrieval and topics were evaluated, including Sentence-BERT, BERTopic, KeyBERT and SciBERT. However, these models have largely been implemented stand-alone and not integrated through an orchestration layer.

Today, the AI-enabled research systems are the following: (1) semantic processing stages are isolated, (2) they do not have adaptive workflow intelligence, (3) they do not have persistent contextual memory, (4) they rely on static retrieval architectures, (5) they provide weak semantic continuity through fragmented conversational interaction, and (6) they have very limited ecosystem orchestration capability.

To overcome those barriers, in this paper, we introduce ResearchLens-CASCO: a Context-Adapt- ive Semantic Cognitive Orchestration Framework for automated literature exploration by means of adaptive semantic cognition and workflow orchestrated AI pipelines. The main contributions of the presenta work can be summarized as follows:

- Proposal of cognitive semantic orchestration framework for the automated exploration of literature.
- Presentation of Dynamic Semantic Context Propagation (DSCP) to maintain contextual coherence over analysis stages.
- Adaptive workflow optimization with ACWO, which adaptively changes the paths of execution according to the quality of retrieval and semantic coherence.
- Persistent semantic reasoning via the Cognitive Research State Memory (CRSM).
- Contextual retrieval and ranking iteratively via the Iterative Research Attention Engine (IRAE). 38
- Combining conversational retrieval and BERTopic based semantic clustering.
- An n8n orchestration based unified workflow-driven cognitive research intelligence architecture that is suitable for a company to build and operate.

II. PROBLEM STATEMENT

General system and algorithm R&D have progressed well in semantic NLP and research automation, but there remain a few essential unsolved challenges for domain-specific literature exploration systems.

A. Non-holistic Semantic Processing

As a result, most existing methods involve performing retrieval, clustering, summarization, and conversational interaction independently and do not maintain contextual continuity across analytical stages. Results generated at a particular stage of the processing are discarded rather than propagated, requiring each processing stage to operate on incomplete context.

B. Predefined Workflow Architectures

Conventional research-analysis systems run at fixed sequence in the workflow without adaptively changing execution routes according to semantic quality or contextual inconclusiveness. A static pipeline continues processing when the retrieval confidence is low or clustering coherence falls off, without recovering from errors and making errors worse and worse along the way.

C. Absence of Persistent Cognitive Memory

However,b/c most conversational retrieval systems lose semantic continuity across interactions, these have_jumped_out_of_lab studies research exploration is fragmented and the analytical reasoning is discontinuous. To the user each session is a clean slate, making it impossible for the system to improve its understanding based on previous interactions.

D. Limitations of Contextual Retrieval

Existing retrieval algorithms are based on the static embedding similarity without evolving semantic refinement or dynamic-aware contextual ranking. Our static ranking ignores the user (researcher)'s evolving search intention, which is obvious through multiple queries.

E. Finite Cognitive Adaptability

None of the existing literature exploration systems has any self-adapting semantic reasoning component that learns and refines the understanding of its context in the course of the workflow execution. These challenges are the combination of 1) adaptive semantic propagation, 2) cognitive memory based persistent centricity, 3) workflow intelligence on process models and 4) multi-perspective centricity based conversational retrieval, and they also establish the rationale behind the design at conceptual level, in the present case ResearchLens-CASCO that amalgamates all presented features into an integrated cognitive research automation system.

III. LITERATURE REVIEW

Semantic document comprehension and smart analysis of literature are emerging research fields in the NLP and AI communities. Sentence-BERT Reimers and Gurevych [1] proposed Sentence-BERT, a method that derives semantically meaningful sentence embeddings which can directly be used for similarity-based ranking and

classification. This seminal work serves as the basis for the embedding layer in ResearchLens-CASCO.

Grootendorst [2] introduced BERTopic, a topic modelling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst being able to find outliers using HDBSCAN. KeyBERT [3] proved that transformer embeddings can also be used for contextual keyword extraction with the help of cosine similarity scoring. They are both embedded as primary features in CASCO Feature Engineering.

Lewis et al. 4] proposed RAG, which borrows elements from retrieval systems such as the inverted index and from generative models and synthesize responses considering context. Beltagy et al. [5] introduced SciBERT, which boosted the performances of NLP tasks over scientific text with a transformer pretrained in a domain-specific manner, and is utilized to abstract domain-specific literature in CASCO.

BERT was first introduced by Devlin et al. (2009) and BM25 (Robertson et al. 2009) to our knowledge.

Although this is a significant advance, most methods attend to single functionalities. So far, few frameworks have provided a collective solution for dynamic semantic context propagation, adaptive workflow orchestration, persistent cognitive memory, iterative retrieval refinement, and conversational semantic continuity all in one research intelligence environment. ResearchLens-CASCO bridges this gap via a context-adaptive cognitive semantic orchestration framework for automated literature investigation.

IV. PROPOSED FRAMEWORK

A. Design Philosophy

Based on the SemmenLens concept, CASCO is constructed on four phylosophic pillars: Semantic Context Continuity, which preserves continuity of context signals at any stage of processing; Adaptive Workflow Iimpact on thes Intelligence, which allows dynamical adjustments of processing (workflow) priority on awareness of processing quality; Persistent Cognitive Memory, which denotes learning and memory accumulation of context inbetween sessions; and Context-Aware Conversational Exploration, which enables conversational exploration through sematic grounding.

B. Overview of the Framework

The proposed CASCO framework consists of the following four tightly coupled SSs**: (i) DSCP, (ii) ACWO, (iii) CRSM and (iv) IRAE. The general process is as follows:**

User Query → Semantic Parsing → Dynamic Context Propagation (DSCP) → Adaptive Workflow Optimization (ACWO) → Contextual Retrieval (IRAE) → Cognitive Memory Update (CRSM) → Conversational Refinement → Analytical Visualization

They are bound together tightly by a shared semantic context vector, which is to say that the output of one module is integrated into the input of the next rather than being a standalone result.

V. SYSTEM ARCHITECTURE

The ResearchLens-CASCO framework is divided into seven modular layers which facilitate the independent development, testing and substitution of the individual modules without affecting the pipeline.

A. Interaction Layer

The frontend interaction layer is built with React.js and offers PDF upload interfaces and conversational research interaction modules a single integrated dashboard visualization, live workflow monitoring, and a module for topic exploration. Events are timestamped with millisecond precision for high-resolution temporal analysis.

B. Semantic Extraction Layer

It includes PDF text extraction, tokenization, preprocessing, lemmatization, sentence splitting, and metadata extraction using spaCy's NLP pipeline. The processed text units (e.g., sentences) of the extracted text are sent to the embedding generation part as normalized inputs.

C. Dynamic Context Layer

The DSCP subsystem dynamically propagates semantic comprehension at all stages of analysis using the context aggregation formula:

$$C\_t = \alpha \cdot E\_t + \beta \cdot T\_t + \gamma \cdot K\_t + \delta \cdot Q\_t + \lambda \cdot H\_t$$

where E_t = semantic embeddings, T_t = topic vectors, K_t = contextual keywords, Q_t = conversational query state, H_t = historical interaction state, and α, β, γ, δ, λ are session-specific adaptive weighting parameters. This formulation to the context vector guarantees that it is holistic from all sources of information and not a unimodal signal.

D. Adaptive Workflow Layer

The ACWO subsystem adapts the workflow execution paths on-the-fly according to retrieval quality, clustering confidence, semantic coherence, and conversational ambiguity. The following determine the best workflow path by:

$$L^* = \arg\max \alpha R + \beta S + \gamma C - \delta A$$

where R = retrieval relevance, S = semantic coherence, C = conversational continuity, and A = ambiguity penalty. When ambiguity arises, ACWO either initiates further retrieval cycles or asks for conversation clarification before advancing.

E. Cognitive Memory Layer

CRSM holds lasting contextual continuity over user sessions with a memory update of exponential moving average form:

$$M\_t = \eta \cdot M\_ \{t-1\} + (1 - \eta) \cdot C\_t$$

where M is the updated state of the cognitive memory, η is the coefficient of memory retention, and C_t is the current vector of semantic context. A higher η favors stability of long-term memory, while a smaller η favors faster adaptation to novel contextual signals. This procedure is parallel to the Temporal Decay Cognitive Memory for (TDCM) in UniMentor-X [Ref], but in this scenario it is adapted to research session continuity instead of learner state modeling.

F. Conversational Retrieval Layer

The conversation subsystem still computes semantic query embedding and contextual retrieval using IRAE, but now dynamically generates prompts, synthesizes relevant passages retrieved from the literature, and iteratively refines the conversation. Answers are based on the retrieved document set and are also influenced by cognitive memory context to help maintain session-level coherence.

G. Visualization and Analytics Layer

This layer enables interactive topic visualization, semantic dashboards, workflow execution analytics, contextual exploration graphs, retrieval quality statistics, allowing the researchers to examine and query each and every step of the analytical pipeline.

VI. METHODOLOGY

A. Representing Research in the Semantic Space

A research paper is modelled as a structured tuple:

$R(t) = \{E(t),T(t),K(t),C(t),M(t)\}$ where E(t) = semantic embedding vector, T(t) = thematic cluster assignment, K(t) = contextual keyword set, C(t) = conversational context state and M(t) = document meta information such as title, authors, publication year, and venue.

B. Semantic Embedding Generation

Sentence-transformer models compute semantic embeddings for each document or query segment:

$E\_i = f\_transformer(D\_i)$ where D_i is the input document or query, and E_i is its representation in a high-dimensional semantic space. Before downstream use, all the embeddings are L2-normalized to guarantee the cosine similarity computations are bounded within [0, 1].

C. Topic Modeling

BERTopic combines semantic embeddings generation, UMAP-based dimensionality reduction, HDBSCAN clustering and class-based TF-IDF topic extraction:

$$T\_i = Cluster(E\_i)$$

Each cluster is labeled with a consistent topic label, calculated from the most discriminative words of each cluster, that can be used as an interpretable summary in the visualization layer.

D. Contextual Keyword Extraction

KeyBERT ranks candidate n-grams based on their cosine similarity with the document embedding to extract contextual keywords:

$$Score(k\_i) = \cos(E\_d, E\_\{k\_i\})$$

Top-ranked keywords are added to the DSCP context vector (K_t), and exposed to the researcher via the dashboard.

E. Iterative Reseach Attention Engine (IRAE)

IRAE adapts the ranking of contextual literature dynamically across retrieval iterations by a multi-source attention score:

$$A\_i = Softmax(W_1 \cdot E\_i + W_2 \cdot C\_i + W_3 \cdot T\_i + W_4 \cdot H\_i)$$

where E_i = embedding similarity to query, C_i = conversational relevance score, T_i = thematic relevance, and H_i = historical interaction importance. Learnable weight matrices W_1–W_4 are optimized on session-level interaction logs. This process of refinement enables IRAE to gradually concentrate on the most contextually pertinent literature along the research session.

F. Contradiction Detection and Resolution

If semantically contradictory claims are made within the retrieved documents, the CASCO pipeline executes a pairwise contradiction scoring function in the embedding space. Documents with high contradiction scores are flagged in the dashboard with explanatory notes allowing the researcher to interrogate rather than be spoon fed a blended synthesis that masks disagreement.

VII. IMPLEMENTATION

The frontend is developed with React.js with the vite build config to enable fast hot-module replacement during development and for optimized production bundles. The backend is built with Node.js and Express.js, which serves RESTful API endpoints that both the frontend and the n8n orchestrator use.

MongoDB contains all persistent artifacts, such as uploaded papers, workflow execution logs, topic cluster assignments, semantic embedding indices, chatbot conversation history, cognitive memory states, and dashboard analytics. A specialized embedding index allows for fast approximate nearest-neighbor retrieval using cosine similarity.

Python NLP micros are full-fied FastAPI endpoints which incorporate BERTopic, KeyBERT, Sentence Transformers and spaCy, orchestrated by n8n workflow automation. Such a microservice breakdown means that single NLP pieces can be upgraded or replaced independently without affecting the rest of the pipeline.

n8n orchestration workflows control the execution order of semantic AI analytical modules, including the ACWO dynamic path selection logic through conditional workflow

branches triggered by quality thresholds on runtime evaluation.

VIII. DATASET AND EXPERIMENTAL SETUP

A. Dataset Collection

For the experimental evaluation, we used research papers taken from arXiv, the Semantic Scholar Open Research Corpus (SSORC), and the PubMed literature subsets. The Entire dataset contain 5000 abstracts of research papers, 2000 full text documents in 15 thematic classes and more than 50000 semantic keywords extracted. Papers were chosen so as to comprise a roughly equal split across the computational, biomedical, and inter-disciplinary research areas.

B. Evaluation Metrics

The framework was tested with Precision@K, Recall@K, F1-Score, Mean Average Precision (MAP), nDCG, Topic Coherence Score (C_v), and Semantic Retrieval Accuracy. Between two of these metrics both retrieval relevance and thematic quality are accounted for.

C. Baseline Systems

ResearchLens-CASCO was contrasted to: TF-IDF retrieval, BM25 retrieval, Vanilla SBERT retrieval, basic spaCy NLP pipelines and static conversational retrieval systems. These baselines span the range of typical methods used in production academic exploration platforms.

IX. EXPERIMENTAL RESULTS

TABLE I. Performance Comparison: ResearchLens-CASCO vs. Traditional Systems

Metric	Traditional Systems	ResearchLens-CASCO
Semantic Retrieval Accuracy	74.5%	92.4% (+17.9 pp)
Contextual Continuity	Low	High (DSCP-driven)
Workflow Adaptivity	Static	Dynamic (ACWO)
Conversational Relevance	Moderate	High (CRSM-enhanced)
Topic Clustering Quality	Moderate	High (BERTopic)
Workflow Automation	Partial	Full (n8n orchestrated)
Research Exploration Efficiency	Moderate	High (IRAE-ranked)

The DSCP subsystem maintained semantic continuity at every level of analysis with no loss of context, and with no loss of context was reported in any of the test sessions. ACWO steadily adapted the workflow execution path for 34% of the queries with the initial retrieval confidence forecast being under the quality threshold, resulting in iterative retrieval cycles that successfully retrieved relevant documents that had been lost by the static baseline. CRSM preserved a stable conversational context across multi-turn dialogue, consequently mitigating repeated clarification requests by 48% in average over stateless retrieval baselines. IRAE enhanced the contextual relevance of retrieval by progressively refining the weights of documents in accordance with session-level attentions, and achieved consistent nDCG gains for all 15 themes.

X. ABLATION STUDY

TABLE II. Ablation Study — Impact of Individual CASCO Components

Configuration	Retrieval Accuracy	Conversational Relevance	Workflow Adaptivity
Without DSCP	81.2%	Moderate	Partial
Without CRSM	84.5%	Moderate	Dynamic
Without ACWO	85.1%	High	Static
Without IRAE	86.3%	High	Dynamic
Full ResearchLens-CASCO	92.4%	Very High	Full

The ablation study shows that Dynamic Semantic Context Propagation yields the largest improvement in accuracy by a single component (11.2 percentage points), confirming that continuity in context is the most important factor in quality of semantic retrieval. CRSM provides the largest increase in conversational relevance, mitigating the deterioration of session-level context. Two of the components (ACWO and IRAE respectively) contribute complementary improvements in workflow robustness and retrieval precision, and all four components show synergetic effects and result in a overall system accuracy of 92.4%.

XI. COMPARATIVE ANALYSIS

TABLE III. Feature Dimension Comparison Across System Architectures

Feature Dimension	Traditional Retrieval	Basic NLP Pipelines	ResearchLens-CASCO
Semantic Continuity	No	Partial	Yes (DSCP)
Adaptive Workflow Intelligence	No	No	Yes (ACWO)

Feature Dimension	Traditional Retrieval	Basic NLP Pipelines	ResearchLens-CASCO
Persistent Cognitive Memory	No	No	Yes (CRSM)
Conversational Context Preservation	Limited	Partial	Advanced (CRSM)
Topic Modeling	Partial	Yes	Advanced (BERTopic)
Workflow Orchestration	Limited	Partial	Full (n8n)
Context-Aware Retrieval	Limited	Moderate	Advanced (IRAE)
Dynamic Semantic Adaptation	No	No	Yes (ACWO + DSCP)
Contradiction Detection	No	No	Yes (embedding-based)

ResearchLens-CASCO is the only approach that seamlessly combines the continuous semantics, adaptive workflow optimization, persistent cognitive memory, iterative attention-based retrieval, and contradiction detection, within one single unified cognitive semantic intelligence model, surpassing all baseline models along each and every evaluated axis.

XII. ORCHESTRATION AND MODULE COMMUNICATION MODEL

The researchlens-casco is a common DSCP context vector and a CRSM memory state through which all inter-module communications are routed; no separate NLP microservices engage in peer-to-peer communications. This design choice guarantees clean separation of concerns, allowing each component to be tested and replaced without affecting the others, which reflects the agent-isolation principle utilized in multi-agent education systems like UniMentor-X.

TABLE IV. Module Communication Contract

Module	DSCP Context Reads	Output Written
BERTopic (Topic Modeling)	E_t, H_t	T_t cluster assignments
KeyBERT (Keyword Extraction)	E_t	K_t keyword scores
IRAE (Retrieval Engine)	E_t, C_t, T_t, H_t	Ranked document list A_i
CRSM (Cognitive Memory)	C_t (current context)	M_t updated memory state
ACWO (Workflow Optimizer)	R, S, C, A metrics	Optimal workflow path L*
Conversational Retrieval	M_t, IRAE output, Q t	Synthesized research response

XIII. DISCUSSION

A. Contributions to the State of the Art

Our 17.9 percent point increase in semantic retrieval accuracy over the best traditional baseline further demonstrates that the combination of DSCP, CRSM, ACWO, and IRAE leads to synergistic improvements beyond what is achievable by each component individually. The ACWO mechanism resolves a form of workflow degradation—processing stall due to unconfident retrieval—that cannot be detected in single-stage evaluation but can be observed in real research sessions. The zero context-loss rate in all DSCP-based test sessions confirms the context propagation design.

2) Limitations.

There are a number of limitations to the present work. Computational cost for generating large-scale embeddings is quadratic in the corpus size for pairwise similarity computation; approximate nearest-neighbor indexing alleviates this somewhat. Workflow execution latency grows with the number of ACWO-mediated remediation cycles, and this may be noticeable by users for large corpora. Multi-column PDF extraction has minor bugs for papers with complex typographic layouts, that sometimes cause malformed sentence boundaries. The current evaluation is based on a dataset curated from three publicly available repositories; a longitudinal deployment study involving live researchers is necessary to establish ecological validity.

C. Ethical Considertions

The framework inherits the semantic biases of pretrained transformer models and the collected literature datasets. Contextual retrieval systems might at times produce semantically believable yet factually false research pairings, an issue further exacerbated when CRSM multi-session strengthens erroneous memory states. Possible biases are retrieval hallucination, dataset representation bias, misinformation dissemination via automated synthesis, and semantic ranking imbalance in favor of highly cited work. Future implementations should combine bias mitigation pipelines, explainable retrieval, human-in-the-loop validation, transparent ranking explanations and dataset diversity balancing to mitigate such risks.

XIV. FUTURE ENHANCEMENTS

Enhancements to be pursued for the ResearchLens-CASCO:

• Support for vector databases (Pinecone, ChromaDB) as a billion-scale approximate nearest-neighbor retriever.

• RAG with multi-hop reasoning chains for question answering on advanced research topics.

• Citation-network intelligence analysis for authority-weighted retrieval and discovery of research gaps.

• Knowledge graph will be generated based on carried out procedure of extracted semantic relations in a literature corpus.

• Multi-agent demos; specialized agents for methodology analysis, result extraction and bias detection that collaborate within the CASCO orchestration layer.

• GPU-accelerated embedding generation pipelines for sub-second latency on corpora of > 100k documents.

• Rule-based ACWO threshold logic is replaced with a policy network, trained on session-level researcher feedback using reinforcement learning based adaptive workflow optimization (part 2).

• Cloud-native distributed orchestration enabling institutional research repositories to scale out horizontally.

• Synchronous Academic API with arXiv/PubMed/Semantic Scholar for Live Updated Corpora.

• Within the CRSM, the access-frequency normalization is embedded for joint institution-researcher fairness and for reducing memory bias against researchers with erratic access patterns.

XV. CONCLUSION

In this paper, we introduced ResearchLens-CASCO, a Context-Adaptive Cognitive-semantics IntelligencE framework to automatically explore the literature at multiple levels of abstraction over flow-based semanticNLP pipelines. The suggested framework combines dynamic propagation of the semantic context, adaptative workflow execution, persistent cognitive memory, iterative refinement of attention, conversational retrieval, semantic clustering, and full workflow orchestration within a single cognitive research intelligence framework.

Three new technical contributions make ResearchLens-CASCO distinct from previous work: DSCP, which makes contextual signals accumulate and propagate throughout every analytical stage rather than lose their strength at stage boundaries; CRSM, which offers persistent memory that grows as the research session goes on; and IRAE, which constructs iteratively refined, multi-source attention-weighted retrieval ranking. The experimental result on 7,000 documents shows 92.4% semantic retrieval accuracy which is 17.9% better than the best traditional baselines and without context-loss event throughout the whole workflow automation in all the tested configurations.

ResearchLens-CASCO serves as a scalable and extensible platform for next generation AI-enabled academic intelligence systems. Future work includes real-world longitudinal deployment with live research cohorts, reinforcement-learning based workflow optimization, multi-agent semantic collaboration, and equity-aware cognitive memory mechanisms to assist researchers with varied access patterns.

REFERENCES.

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP, 2019.

[2] M. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," arXiv preprint arXiv:2203.05794, 2022.

[3] M. Grootendorst, "KeyBERT: Minimal Keyword Extraction with BERT," 2020.

[4] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.

[5] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in Proc. EMNLP, 2019.

[6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[7] M. Honnibal and I. Montani, "spaCy 2: Natural Language Understanding with Bloom Embeddings," 2017.

[8] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020.

[9] React Development Team, "React Documentation," 2024. [Online]. Available: https://react.dev

[10] MongoDB Inc., "MongoDB Documentation," 2024. [Online]. Available: https://www.mongodb.com/docs

[11] n8n GmbH, "n8n Workflow Automation Documentation," 2024. [Online]. Available: https://docs.n8n.io

[12] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009.

[13] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv preprint arXiv:1802.03426, 2018.

[14] R. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in Proc. PAKDD, 2013.

[15] P. Schwab and W. Karlen, "CXPlain: Causal Explanations for Model Interpretation under Uncertainty," in Proc. NeurIPS, 2019.

