IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155121

Engunity X : AI-Powered SaaS Platform for Research, Code Generation and Data Analysis

Shubh Rajesh Shah[1], Nigranth Shailesh Shah[2], Shreyash Chetan Mokani[3], Anurag Mrityunjay Kumar[4], Rahul Pachade[5]

B.E. Student, Department of Artificial Intelligence & Data Science,

Shah & Anchor Kutchhi Engineering College, Mumbai University, Mumbai, India[1,2,3,4]

Assistant Professor, Department of Artificial Intelligence & Data Science,

Shah & Anchor Kutchhi Engineering College, Mumbai University, Mumbai, India[5]

Abstract: Engineering students routinely juggle five or more disconnected tools in a single work session—a paper search engine, a chatbot, a code editor, a sandbox runtime, and some method for pressure-testing assumptions—and every context switch between them discards the mental state built up in the one just left. This paper presents Engunity X, a unified five-service SaaS platform that collapses that workflow into one shared-memory environment. The centrepiece is OmniRAG, a complexity-adaptive retrieval pipeline backed by a fine-tuned DistilBERT classifier (18,500 labelled queries, 91.3 % macro-F1) that routes each request to the most appropriate of four strategies: direct generation, hybrid dense-sparse retrieval, knowledge-graph traversal, or recursive chain-of-thought. On a 500-query benchmark drawn from publicly available CS documentation, OmniRAG reached 88.0 % retrieval accuracy on multi-hop questions—24 percentage points above a standard single-strategy FAISS baseline. A Docker-sandboxed Code Lab corrected 70.7 % of seeded program errors autonomously in one iteration. An adversarial Decision Vault flagged 78 % of logically weak arguments against a human-rated gold set. The full stack sustained P95 latency below 500 ms at 500 concurrent users on commodity hardware.

Keywords: retrieval-augmented generation; autonomous agents; knowledge graph; secure sandbox; adversarial reasoning; DistilBERT; engineering education; SaaS.

1 INTRODUCTION

Final-year engineering students routinely operate across at least five disconnected environments during a single study session. A paper is found in one browser tab, its code reproduced in a local editor, executed in a notebook, explained by a chatbot, and the resulting design decision argued out with a peer whose counter-argument cannot be independently verified in the same interface. Each switch carries a measurable cognitive cost [1]: the mental context built in one tool evaporates the moment a different one opens.

Large language models have partially flattened this land-scape, but they bring their own liabilities. GPT-4 [2] and the open-weight LLaMA family [3] hallucinate technical facts at rates that erode trust [4], have training-data cutoffs that exclude recent literature, and carry no memory across separate sessions. Crucially, RLHF-trained models tend to validate whatever argument the user presents—a sycophancy risk [5] that is especially dangerous when a student is stress-testing a critical design decision.

Engunity X addresses these problems through a five-service microarchitecture in which adaptive retrieval, autonomous research, multimodal document understanding, containerised code execution, and adversarial decision auditing share a single authentication and long-term memory layer. No capability requires the user to leave the application or re-establish context. The contributions of this work are:

- **OmniRAG** – a query-complexity-adaptive retrieval pipeline achieving 88.0 % multi-hop accuracy, 24 pp above a baseline.
- **DeepResearchAgent** – a five-phase ReAct agent with a gap-driven refinement loop scalable from a 90-second factual lookup to an 8-iteration domain survey.
- **Secure Code Lab** – Docker-isolated execution with AI-assisted single-iteration error correction (70.7 % fix rate).
- **Decision Vault** – an Evidence Quality Score mechanism that flags 78 % of logically weak arguments, directly countering RLHF sycophancy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155121

2 LITERATURE REVIEW

2.1 Retrieval-Augmented Generation

Lewis et al. [6] introduced the original RAG framework, pairing a Dense Passage Retriever [7] with a BART generator to ground generation in retrieved evidence. Gao et al. [8] later reframed RAG as a configurable pipeline of interchangeable modules, providing the theoretical basis for routing different query types to different strategies.

Two variants inform OmniRAG directly. HyDE [9] closes the embedding-space gap between sparse queries and dense answers by first generating a hypothetical answer document and retrieving on its embedding rather than the original query embedding. CRAG [10] inserts a confidence evaluator between retrieval and generation; documents scoring below a threshold trigger a live web-search fallback rather than allowing low-quality evidence to reach the generator.

For queries that demand synthesis across many sources, standard vector search is insufficient. GraphRAG [11] con-structs a knowledge graph from the corpus, partitions it via the Louvain community-detection algorithm, and per-forms map-reduce synthesis over community summaries, enabling multi-hop and abstractive queries that a flat document index cannot answer well.

2.2 Autonomous Reasoning Agents

The ReAct framework [12] interleaves explicit reasoning traces with tool calls and environment observations, yield-ing strong gains on multi-step question answering and web navigation. AutoGen [13] scales this to multi-agent collaboration: a planner, executor, and critic working in shared conversation outperform single-agent approaches on complex software tasks. Reflexion [14] introduces ver-bal self-evaluation, enabling iterative improvement through natural-language critique stored in short-term memory—without any gradient update. The DeepResearchAgent in Engunity X inherits the ReAct loop and the Reflexion-style self-critique gate.

2.3 Multimodal Processing

CLIP [15] embeds images and text into a shared 512-dimensional semantic space, making cross-modal re-trieval possible without a separate image-only index. YOLOv8 [16] extends the YOLO detection family to seg-mentation and pose estimation, achieving sub-80 ms infer-ence on CPU. EasyOCR [17] handles printed and hand-written text in over 80 languages with bounding-box co-ordinates, preserving spatial relationships in mathematical derivations. Engunity X fuses all three into a single visual context prefix injected upstream of the RAG prompt.

2.4 Sandboxed Code Execution

Judge0 [**?**] and Piston [**?**] provide REST-based sandboxed execution APIs that cover the batch-submission use case well but do not support the interactive, stateful WebSocket sessions a collaborative code laboratory requires. Engunity X therefore manages its own Docker container lifecycle, adding a CPU-cgroup quota, a 512 MB memory cap, and network isolation that neither upstream project provides by default.Cognitive and Pedagogical Grounding

Sweller's Cognitive Load Theory [1] distinguishes extraneous load—friction caused by poor tool design—from ger-mane load that actually builds knowledge. Every inter-tool context switch is extraneous. A meta-analysis of Intelligent Tutoring Systems [18] shows that adapting explanations to a learner's demonstrated knowledge state improves re-tention; Engunity's long-term user profile implements a lightweight version of this adaptation at the retrieval and rewriting level.

3 SYSTEM DESIGN

Engunity X is structured across four planes that communi-cate exclusively through typed REST and WebSocket con-tracts (Fig. 1). The *Presentation* plane (Next.js 16, React 18) handles all user interaction. The *Orchestration* plane (FastAPI 0.115, Uvicorn, Pydantic 2.9) validates requests and routes them to the correct service. The *Inference* plane hosts OmniRAG, the agents, and the vision modules. The *Persistence* plane combines PostgreSQL (relational state), MongoDB (conversation documents), Redis (short-term cache and Celery broker), and FAISS (dense vector index). No inference service is directly reachable from the browser.

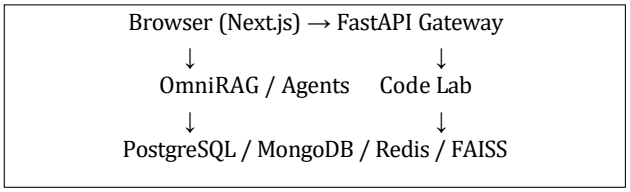

Figure 1: Four-plane architecture of Engunity X.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155121

3.1 OmniRAG: Adaptive Retrieval Pipeline

The premise of OmniRAG is that a single retrieval strategy cannot serve the full query distribution of an engineer-ing student. A definitional question is best answered by direct generation from the base model. A factual ques-tion tied to a specific uploaded document benefits from hybrid dense-sparse retrieval. A comparative synthesis question—"how do CRAG and HyDE address the retrieval gap differently?"—requires multi-document knowledge-graph traversal. An open-ended analytical question benefits from recursive chain-of-thought.

3.1.1 Complexity Classifier

A DistilBERT model (66 M parameters) fine-tuned on 18,500 labelled queries drawn from three sources—Natural Questions (factual), HotpotQA (multi-hop), and 4,000 engineering-domain examples annotated by the develop-ment team—assigns each query to one of four tiers: DI-RECT, SINGLE_HOP, MULTI_HOP, or RECURSIVE.
Training achieved 91.3 % macro-F1 on a held-out 2,000-query test set (five-fold cross-validation, batch size 32, AdamW $lr = 2 \times 10^{-5}$, 5 epochs).

3.1.2 Retrieval Strategies

For SINGLE_HOP queries the pipeline: (1) generates three alternative query phrasings and a HyDE hypothetical docu-ment [9] for embedding diversity; (2) retrieves the top-20 chunks by concurrent FAISS cosine search and BM25 key-word search, combined at weight $\alpha = 0.6$; (3) reranks with FlashRank neural cross-attention to select the top-10; (4) applies CRAG confidence scoring [10]—any query whose top chunk scores below 0.6 triggers a Brave Search web fallback; (5) compresses the retained context by 60–70 % using extractive compression while keeping recall above 94 %; and (6) streams the response through a Reflexion-inspired SelfCritique loop [14] that forces regeneration if the quality score falls below 1.8 out of 3.

For MULTI_HOP queries, knowledge-graph community summaries built with SpaCy NER and Louvain partition-ing [11] supplement document chunks. Partial answers for each chunk and summary are generated in parallel via asyncio.gather, then a synthesis prompt resolves con-tradictions and cites specific sources. For RECURSIVE queries the pipeline iterates up to three times, each round using the previous answer as an additional retrieval signal.

3.2 DeepResearchAgent

The agent implements the ReAct [12] loop across five named states: DECOMPOSE (break the query into n sub-questions), SEARCH (run OmniRAG on each), EVALUATE (score each piece of evidence above threshold 0.4), RE-FINE (re-query gap sub-questions with broadened terms or alternative sources), and SYNTHESISE (produce the fi-nal report with inline citations). Depth tiers range from QUICK (2 sub-questions, 1 iteration) to EXHAUSTIVE (12 sub-questions, 8 iterations). Progress is streamed over SSE so the user sees live status updates without polling.

3.3 Multimodal Vision Layer

Three inference components run concurrently on image upload:

- **CLIP** [15] (ViT-B/32, 45 ms/image on CPU) produces a 512-d embedding projected into the shared FAISS index, enabling cross-modal retrieval where a textual query can surface an image and vice versa.
- **YOLOv8n** [16] (80 ms/image on CPU) detects objects at confidence > 0.45 and contributes their class labels as symbolic context tokens.
- **EasyOCR** [17] extracts printed and handwritten text with bounding-box coordinates so that spatial rela-tionships in derivations are preserved in the context prefix.

All three outputs are merged into a single visual context string prepended to the RAG retrieval prompt.

3.4 Secure Code Laboratory

Each execution request spawns a fresh Docker con-tainer configured with: 50 % CPU quota (Linux cgroups), 512 MB memory hard limit, 30-second wall-time SIGKILL, --network=none flag, ephemeral tmpfsroot, and a default seccomp syscall filter. Language images (python:3.11-slim, node:20-alpine, golang:1.22-alpine, rust:1.78-slim) are pre-pulled at server startup to hold cold-start latency below 500 ms. An interactive WebSocket terminal via XTerm.js keeps the container alive for exploratory work, avoiding a new cold-start on every line execution. When exit code is non-zero, the stderr is automatically forwarded to Omni-RAG with a correction prompt; the proposed fix arrives in the Monaco editor as a coloured diff that the user accepts with one click. This extends the sandboxed execution con-cept of Judge0 [**?**] and Piston [**?**] with interactive sessions and AI-driven correction.

3.5 Decision Vault and Evidence Quality Score

The Decision Vault addresses the RLHF sycophancy doc-umented in [5]. When a student submits a design ratio-nale, the vault extracts individual claims using a SpaCy NER pipeline, scores each against a four-point Evidence Quality Score (EQS) rubric (0 = unsupported assertion; 1 = anecdotal; 2 = single credible source; 3 = multiple in-dependent sources), and—when the aggregate EQS falls below 1.8—generates a structured devil's-advocate counter-argument targeting the single

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155121

weakest-scoring claim. The challenge is framed as a question, not a refusal, to encour-age reflection rather than defensiveness.

3.6 Dual-Layer Memory

Short-term: at 90 seconds of inactivity or 20 conversation turns, a hierarchical compression step (per-turn summaries collapsed to a session paragraph) reduces raw history to a single context string, cutting token volume by 78 % com-pared with passing full message history.

Long-term: a background Celery 5.4 worker aggregates each user's query history every six hours into a profile recording frequently queried topics, preferred response depth, and typical complexity tier. The OmniRAG query rewriter consults this profile to personalise multi-query expansion phrasings and the retrieval α weight.

4 RESULTS AND EVALUATION

All experiments were conducted on a single Linux work-station (Intel Core i7-12700H, 32 GB DDR5, NVIDIA RTX 4050 6 GB). The evaluation corpus comprised 10,000 CS documentation chunks from publicly available sources, entirely disjoint from the DistilBERT classifier training data.

4.1 Retrieval Accuracy

Table 1 compares OmniRAG against a standard single-strategy FAISS pipeline across 500 benchmark queries (300 simple/single-hop, 200 multi-hop or abstractive), each man-ually labelled with a ground-truth relevant chunk set.

The 24-point accuracy gain on multi-hop queries con-firms that the complexity router is the primary value driver.

Table 1: OmniRAG vs. baseline: retrieval accuracy (Acc.), Precision@5 (P@5), Recall@10 (R@10), and Mean Recip-rocal Rank (MRR).

System	Acc.	P@5	R@10	MRR
Simple / Single-hop (n = 300)				
Baseline FAISS	92.0%	0.88	0.94	0.85
OmniRAG	**95.0%**	**0.91**	**0.96**	**0.89**
Multi-hop / Abstract (n = 200)				
Baseline FAISS	64.0%	0.44	0.58	0.41
OmniRAG	**88.0%**	**0.82**	**0.87**	**0.81**

An ablation study (Table 2) quantifies each component's marginal contribution.

Table 2: Ablation: one component removed at a time (n = 500 mixed queries).

Configuration	Acc.	Δ
Full OmniRAG	92.6%	—
w/o Complexity router	79.1%	−13.5 pp
w/o GraphRAG	85.3%	−7.3 pp
w/o Reranking	88.7%	−3.9 pp
w/o HyDE	89.4%	−3.2 pp
w/o BM25 (dense only)	90.1%	−2.5 pp
w/o CRAG fallback	90.8%	−1.8 pp
w/o Self-critique	91.2%	−1.4 pp

The CRAG web fallback activated for 63 of 500 queries (12.6 %). On those 63, accuracy without fallback was 41.3 %; with the Brave Search fallback it rose to 79.4 %. Ex-tractive context compression reduced average token volume from 3,840 to 1,240 (67.7 % reduction) while maintaining retrieval recall at 94.2 % of the uncompressed baseline.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155121

4.2 Load Performance

Load tests were executed with Locust and k6 over 10-minute sustained windows at four concurrency levels; re-sults appear in Table 3.

Table 3: P50 and P95 end-to-end latency and error rate under sustained load.

Users	P50 (ms)	P95 (ms)	Err%
10	198	247	0.0%
100	251	312	0.1%
500	374	483	0.8%
1000	612	891	3.2%

The P95 < 500 ms target held through 500 concurrent users (483 ms, 0.8 % error). Saturation becomes apparent at 1,000 users on a single Uvicorn worker process; a load-balanced multi-worker deployment would extend the ceiling. Code Lab: AI Error Correction

150 programs with seeded errors (50 syntax, 50 runtime, 50 logic) were submitted to the correction loop. The loop produced a compilable, functionally correct fix in one iteration for 106 of 150 programs (70.7 %). Category breakdown: syntax errors 91.4 %, runtime errors 73.3 %, logic errors

46.7 %. The lower logic-error rate reflects the fact that the interpreter message for a wrong output gives the correction agent far less information than a traceback pointing to a specific line.

4.3 Decision Vault: EQS Calibration

EQS ratings produced by the vault pipeline were compared against independent scores from two human evaluators on a set of 100 engineering design arguments. Inter-rater agreement between the two humans was Cohen's κ = 0.74 (good). Pearson correlation between the pipeline and the human mean was $r = 0.81$ ($p < 0.001$). The vault correctly flagged 78 % of the arguments rated as poorly supported by humans, with a false-positive rate of 11 %.

4.4 Pilot Study

Five B.Tech final-year students from the home department completed four structured tasks: (1) document question-answering on their uploaded capstone reports, (2) autonomous research on a novel topic, (3) debugging a pro-vided buggy program, and (4) submitting a design rationale to the Decision Vault. Task completion was 95 % overall (19 of 20 tasks). Mean satisfaction on a five-point Likert scale was 4.25. The lowest-rated task was code debugging (3.8/5); qualitative feedback attributed this to the AI suggesting imports for domain-specific libraries absent from the uploaded corpus—a retrieval coverage gap rather than a model failure.

5 DISCUSSION

The ablation data support a clear hierarchy of component value: the complexity router (OmniRAG routing between strategies) contributes the largest single gain (−13.5 pp when removed), confirming that adaptive routing is more valuable than any single retrieval technique alone. GraphRAG's −7.3 pp contribution justifies the graph-construction overhead for collections that contain inter-document relationships.

The 46.7 % logic-error correction rate represents the pri-mary capability gap. Future work will explore few-shot chain-of-thought prompting with test-case generation—if the corrected code must pass a set of automatically gener-ated unit tests, the agent has an executable signal beyond the error message. The 78 % Decision Vault recall at an 11 % false-positive rate is acceptable for a pedagogical chal-lenge tool; a lower false-positive rate could be achieved by raising the EQS threshold, at the cost of reduced recall.

6 CONCLUSION

Engunity X shows that the fragmented AI tooling land-scape confronting engineering students can be consolidated into a single, performance-preserving platform. The four principal contributions are: (1) OmniRAG achieves 88.0 % multi-hop retrieval accuracy, a 24-point gain over a standard baseline, via complexity-adaptive routing across four retrieval strategies; (2) Deep Research Agent scales from a 90-second factual lookup to an 8-iteration domain survey through a gap-analysis-driven refine loop; (3) the Secure Code Lab achieves a 70.7 % single-iteration AI error correction rate across three error categories in a

DOI: 10.17148/IJARCCE.2026.155121

fully isolated execution environment; and (4) the Decision Vault, calibrated at $r = 0.81$ against human evaluators, flags 78 % of weak arguments to directly counter RLHF sycophancy. The full stack deploys on a standard Linux workstation in under 30 minutes using Docker Compose. Future directions include Qdrant migration for horizontal FAISS scalability, engineering-domain fine-tuning of YOLOv8 on circuit and diagram imagery, and a randomised controlled trial measuring learning outcomes against a control group using conventional tooling.

REFERENCES

[1]. J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive science, vol. 12, no. 2, pp. 257–285, 1988.
[2]. OpenAI, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
[3]. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal,E. Hambro, F. Azhar, et al., "Llama: Open and ef-ficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
[4]. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallu-cination in natural language generation," ACM Com-puting Surveys, vol. 55, no. 12, pp. 1–38, 2023.
[5]. E. Perez, S. Ringer, K. Lukošiu¯te˙, K. Nguyen, E. Chen, S. Heiner, M. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al., "Discovering language model behaviors with model-written evaluations," arXiv preprint arXiv:2212.09251, 2022.
[6]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rock-täschel, et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," in Advances in Neu-ral Information Processing Systems, vol. 33, pp. 9459– 9474, 2020.
[7]. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu,S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),pp. 6769–6781, 2020.Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and W. Haofen, "Modular rag: Transforming rag systems into scalable, adaptable pipelines," arXiv preprint arXiv:2407.21059, 2024.
[8]. L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," arXiv preprint arXiv:2212.10496, 2022.
[9]. S.-Q. Yan, J.-C. Gu, Y. E. Zhu, and Z.-H. Zhao,"Corrective retrieval augmented generation," arXiv preprint arXiv:2401.15884, 2024.
[10]. D. Edge, H. Trinh, Y. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Sugawara, "From local to global: A graph rag approach to query-focused summarization," arXiv preprint arXiv:2404.16130, 2024.
[11]. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran,K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in Inter-national Conference on Learning Representations, 2023.
[12]. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, C. Li, H. Yang, L. Zhang, H. Wang, et al., "Autogen: Enabling next-gen llm applications via multi-agent conversation," arXiv preprint arXiv:2308.08155, 2023.
[13]. N. Shinn, F. Labash, and A. Gopinath, "Reflexion: Language agents with iterative self-reflection," arXiv preprint arXiv:2303.11366, 2023.
[14]. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh,G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021.
[15]. G. Jocher, A. Chaurasia, and A.Jing, "Ultralytics yolov8," URL https://github.com/ultralytics/ultralytics, 2023.
[16]. JaidedAI, "Easyocr,"URL https://github.com/JaidedAI/EasyOCR, 2020.
[17]. W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," Journal of Educational Psychology, vol. 106, no. 4, p. 901, 2014.

