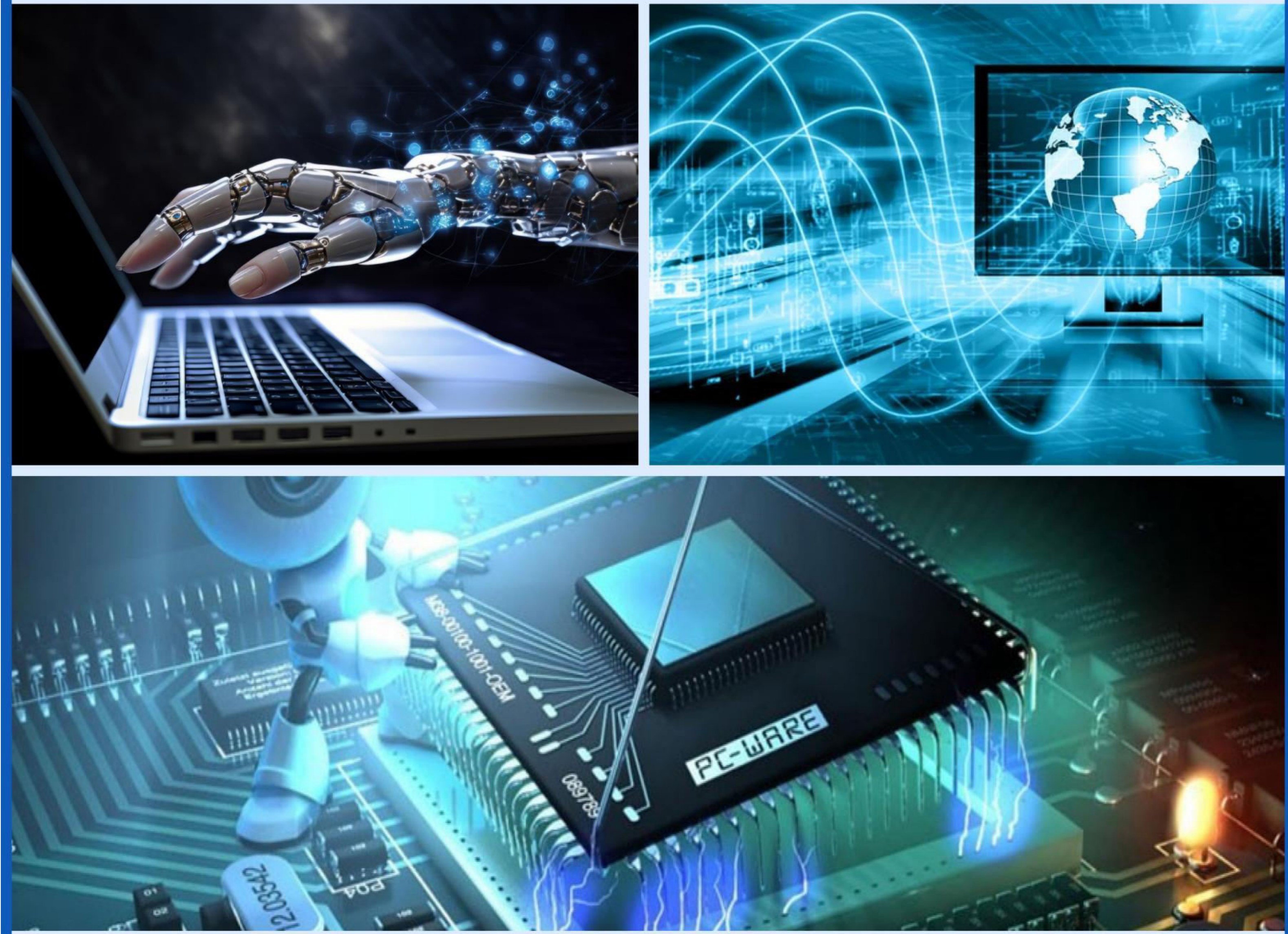
ISSN: 2320-0081

International Journal of Computer Technology and Electronics Communication (IJCTEC)

(A Biannual, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Volume 8, Issue 5, September - October 2025

www.ijctece.com ijctece@gmail.com +91 9710583466 +91 9710583466

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805001

RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines

Dr. Rashmiranjan Pradhan

AI, Gen AI, Agentic AI Innovation leader at IBM, Bangalore, Karnataka, India

ABSTRACT: Retrieval-Augmented Generation (RAG) has emerged as a cornerstone for building context-aware and factual Large Language Model (LLM) applications. However, evaluating the performance of these complex pipelines remains a significant challenge. Existing evaluation frameworks often focus on a narrow set of metrics, failing to provide a holistic view of a system's accuracy, reliability, and practical usability. This paper introduces RAGEvalX, an extended, multi-faceted evaluation framework designed to address this gap. RAGEvalX systematically measures four crucial dimensions: (1) Core RAG Accuracy, including faithfulness and relevancy; (2) Context Integrity, assessing the quality and utilization of retrieved information; (3) Robustness against common input perturbations; and (4) Practical Statistics for operational monitoring. We provide a detailed methodology for implementing the framework, complete with code snippets and guidance on LLM selection for evaluation tasks. Through case studies in healthcare, finance, and legal sectors, we demonstrate how RAGEvalX provides actionable insights for optimizing RAG pipelines. Our framework offers a standardized, comprehensive, and implementable approach to ensure RAG systems are not only accurate but also reliable and ready for real-world deployment.

KEYWORDS: “Retrieval-Augmented Generation,” " RAG,” " Large Language Models,” " LLM,” " Evaluation Metrics,” " AI Robustness,” " Natural Language Processing,” " IEEE Standards."

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) is a powerful paradigm that enhances the capabilities of Large Language Models (LLMs) by grounding them in external, up-to-date knowledge sources. By retrieving relevant documents before generation, RAG pipelines mitigate hallucinations, improve factual accuracy, and enable domain-specific applications. As enterprises increasingly deploy RAG systems for critical functions—from clinical decision support in healthcare to financial analysis—the need for rigorous, comprehensive evaluation has become paramount.

However, the evaluation of RAG systems is non-trivial. The performance is a function of its interconnected components: the retriever, the re-ranker, and the generator. A failure in any component can degrade the final output. Early evaluation efforts focused on component-level metrics or end-to-end answer quality using metrics like Faithfulness and Answer Relevancy. While foundational, these metrics do not capture the full picture. For instance, a system might produce a faithful and relevant answer but fail to utilize the most critical piece of retrieved context, or it may break down when faced with a minor typo in the user's query.

This paper identifies a critical research gap: the absence of a unified framework that integrates core performance metrics with context quality, robustness testing, and practical operational statistics. Existing tools and frameworks often operate in silos, forcing developers to stitch together multiple solutions for a complete evaluation. To address this, we propose **RAGEvalX**, an extended framework designed for holistic RAG pipeline assessment.

The contributions of this work are threefold:

1. We introduce a novel, four-dimensional framework that consolidates Core Accuracy, Context Integrity, Robustness, and Practical Statistics.
2. We provide a detailed, implementable guide with pseudo-code, outlining how to measure each metric using modern LLMs and libraries.

3. We demonstrate the framework's utility through diverse, industry-specific case studies, offering insights into model selection and pipeline optimization.

II. RELATED WORK

The evaluation of RAG systems has evolved rapidly. Initial approaches borrowed from question-answering (QA) benchmarks, using exact match (EM) and F1 scores. However, these are often too rigid for the nuanced, generative nature of LLMs.

More recent work has focused on LLM-as-a-judge methodologies, where a powerful model evaluates the quality of a RAG pipeline's output. This has given rise to several key metrics and frameworks.

RAGAS is a prominent framework that offers component-wise evaluation without relying on ground-truth human annotations. It introduced metrics like Faithfulness, Answer Relevancy, Context Precision, and Context Recall. RAGAS is highly effective for assessing the core functionality of the retriever and generator.

TruLens provides tools for tracking and evaluating LLM applications, including RAG. It focuses on a "triad" of evaluations: ground-truth agreement, summary agreement (for faithfulness), and relevance to the prompt.

ARES focuses on improving the efficiency and accuracy of RAG evaluation by using fine-tuned small models as evaluators, reducing the cost and latency associated with using large proprietary models like GPT-4.
While these frameworks provide an excellent foundation, they have limitations. RAGAS's Context Recall requires a ground-truth answer, which is often unavailable. Robustness testing is not a primary focus, and practical statistics like answer length or context count, which are vital for production monitoring, are typically handled separately. RAGEvalX builds upon the principles established by these frameworks, extending them into a more comprehensive and practical structure aimed at production-readiness.

III. METHODOLOGY

The RAGEvalX framework is designed to be modular and progressive, allowing teams to adopt components as their RAG systems mature. The overall workflow is depicted in Fig. 1.

Fig. 1. The RAGEvalX Evaluation Workflow. A user query initiates the RAG process. RAGEvalX intercepts the inputs and outputs at each stage (Retrieval, Generation) to compute metrics across its four core modules.
The framework consists of four evaluation modules, which can be implemented as a step-by-step process.

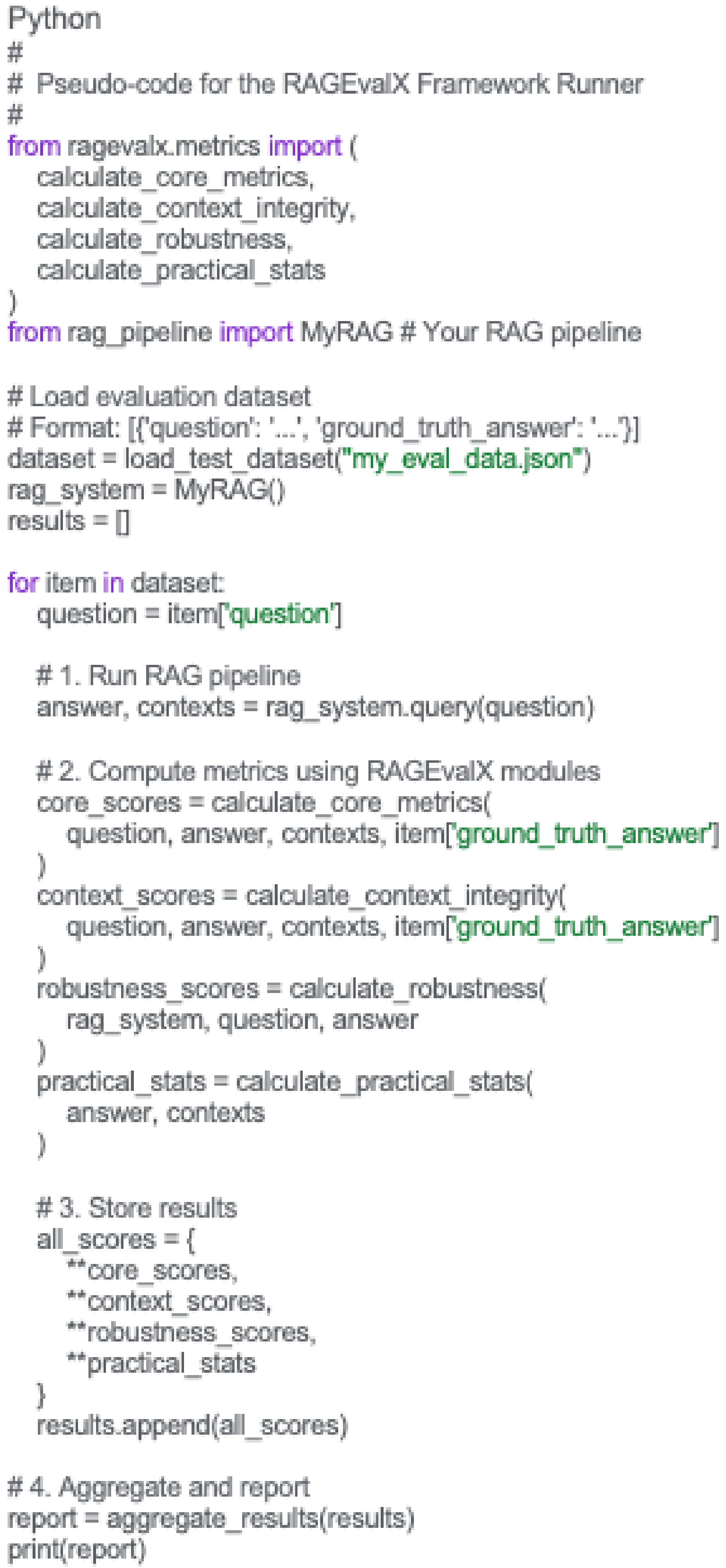
A. RAGEvalX Design

The framework is built around a synthetic test dataset containing (question, ground_truth_answer, ground_truth_context). For many metrics, only the question is needed.
Step 1: Core Metric Generation. For each question, run the RAG pipeline to generate (answer, retrieved_contexts).
Step 2: Module Execution. Pass the (question, answer, retrieved_contexts) tuple, along with any ground truth data, to the four RAGEvalX modules for metric computation.
Step 3: Aggregation and Analysis. Aggregate the scores for each metric across the entire dataset to produce a final report.

B. Implementation Guide

We use Python with libraries like langchain, llama-index, and ragas. The evaluation itself relies on LLM-as-a-judge calls, typically to a powerful model like GPT-4 or Claude 3 Opus.

Below is a Pythonic pseudo-code representation of the RAGEvalX runner.



```
Python
#
# Pseudo-code for the RAGEvalX Framework Runner
#
from ragevalx.metrics import (
    calculate_core_metrics,
    calculate_context_integrity,
    calculate_robustness,
    calculate_practical_stats
)
from rag_pipeline import MyRAG # Your RAG pipeline

# Load evaluation dataset
# Format: [{'question': '...', 'ground_truth_answer': '...'}]
dataset = load_test_dataset("my_eval_data.json")
rag_system = MyRAG()
results = []

for item in dataset:
    question = item['question']

    # 1. Run RAG pipeline
    answer, contexts = rag_system.query(question)

    # 2. Compute metrics using RAGEvalX modules
    core_scores = calculate_core_metrics(
        question, answer, contexts, item['ground_truth_answer']
    )
    context_scores = calculate_context_integrity(
        question, answer, contexts, item['ground_truth_answer']
    )
    robustness_scores = calculate_robustness(
        rag_system, question, answer
    )
    practical_stats = calculate_practical_stats(
        answer, contexts
    )

    # 3. Store results
    all_scores = {
        **core_scores,
        **context_scores,
        **robustness_scores,
        **practical_stats
    }
    results.append(all_scores)

# 4. Aggregate and report
report = aggregate_results(results)
print(report)
```

DOI: 10.15680/IJCTECE.2025.0805001

IV. METRICS & EVALUATION

This section provides a detailed definition of each metric within the RAGEvalX framework, organized by module.

Module 1: Core RAG Metrics
These are fundamental metrics assessing the end-to-end quality of the generated answer.

Faithfulness: Measures if the generated answer is factually consistent with and grounded in the retrieved contexts. An unfaithful answer is a hallucination.

- Implementation: An evaluator LLM is prompted to cross-reference every statement in the answer against the provided contexts. **GPT-4** and **Claude 3 Opus** are highly effective due to their strong reasoning capabilities.

Answer Relevancy: Measures how well the answer addresses the user's question. An answer can be faithful but irrelevant if it doesn't satisfy the user's intent.

- Implementation: An evaluator LLM scores the relevance on a scale of 1-5, considering the directness and completeness of the response to the query.

Context Relevance: Measures the signal-to-noise ratio of the retrieved contexts. Are the retrieved chunks pertinent to the question?

- Implementation: For each (question, context) pair, an LLM evaluates if the context is necessary to answer the question. The final score is the ratio of relevant chunks to the total number of chunks.

Context Precision: A retriever-focused metric that complements Context Relevance. It evaluates whether the ordering of retrieved chunks is optimal, with the most relevant ones ranked highest.

- Implementation: Weighted evaluation where higher-ranked chunks contribute more to the final score.

Module 2: Context Integrity
This module goes deeper than core metrics to assess how well the retrieved context is formed and utilized.

- **Context Precision (No Reference):** A variant of Context Precision that measures the factual consistency between the question and the retrieved context. It is a proxy for retriever quality when no ground truth is available.
- **Context Utilization:** Measures the extent to which the generated answer uses the provided contexts. A low score indicates the LLM is ignoring the retrieved information, potentially relying on its parametric knowledge.
  Implementation: An evaluator LLM identifies which parts of the context were used to generate the answer. The score is the ratio of utilized context to the total context.
- **Context Recall:** Measures the retriever's ability to fetch all necessary information required to answer the question, based on a ground-truth answer.
  Implementation: The ground-truth answer is parsed into statements. For each statement, an LLM checks if the retrieved contexts contain the information to support it.
- **Context Entity Recall:** A fine-grained version of Context Recall, crucial for domains like finance and legal. It measures the percentage of key entities (e.g., names, dates, figures) from the ground-truth context that are present in the retrieved context.

Module 3: Robustness Metrics
This module evaluates the pipeline's resilience to common real-world challenges.

- **Noise Sensitivity:** Measures how much the output quality degrades when small, non-semantic perturbations (e.g., typos, extra whitespace) are added to the input question.

Implementation: Generate a perturbed version of the question. Run both the original and perturbed questions through the RAG pipeline. Use an evaluator LLM to measure the semantic similarity or factual consistency between the two answers. A high score indicates high robustness.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805001

- **Response Relevancy (Negative Examples):** Tests the system's ability to gracefully handle out-of-scope or irrelevant questions. The system should ideally respond that it cannot answer, rather than retrieving irrelevant documents and generating a poor answer.

Module 4: Practical Statistics
These simple but vital metrics are essential for production monitoring and cost management.

- **Answer Length:** The word or token count of the generated answer.
- **Context Count:** The number of retrieved context chunks.

Table I: RAGEvalX Metric Summary

Module	Metric	Description	Recommended Evaluator LLM
Core	Faithfulness	Is the answer grounded in the context?	GPT-4, Claude 3 Opus
	Answer Relevancy	Does the answer address the question?	Mistral-Large, Llama-3-70B
	Context Relevance	Are the retrieved chunks relevant?	Mistral-Large, Llama-3-70B
Context	Context Utilization	How much of the context was used in the answer?	GPT-4, Claude 3 Opus
	Context Entity Recall	Were key entities from the ground truth retrieved?	GPT-4
Robustness	Noise Sensitivity	How does the answer change with input typos?	Llama-3-70B, GPT-4
Practical	Answer Length	Word/token count of the answer.	N/A (Direct Calculation)
	Context Count	Number of chunks retrieved.	N/A (Direct Calculation)

V. EXPERIMENTS & CASE STUDIES

We applied RAGEvalX to three simulated industry-specific RAG pipelines.

A. Healthcare: Clinical Query System

A RAG system designed to answer clinician questions based on a knowledge base of medical research papers (e.g., PubMed abstracts).
Objective: Maximize factual accuracy and ensure all key clinical details are retrieved.

RAGEvalX Application:
High-Stakes Metrics: Faithfulness and Context Entity Recall were prioritized. A hallucinated answer or a missed dosage entity could have severe consequences.

Findings: The initial pipeline using a basic vector search scored 0.92 on Faithfulness but only 0.65 on Context Entity Recall. By implementing a hybrid search (semantic + keyword), we improved Context Entity Recall to 0.88 with a negligible drop in Faithfulness.

LLM Guidance: GPT-4 was found to be the most reliable evaluator for faithfulness, consistently catching subtle factual inconsistencies that other models missed.

Finance: Financial Report Analysis
A RAG system that ingests quarterly financial reports (10-K filings) to answer analyst questions.
Objective: Provide precise numerical answers and handle complex, jargon-heavy queries.

RAGEvalX Application:
High-Stakes Metrics: Noise Sensitivity and Context Utilization. Analysts may make typos, and the system must rely on the retrieved report figures, not its parametric knowledge.

Findings: The system was highly sensitive to acronym variations (e.g., "EBITDA" vs. "earnings before interest..."). We implemented a query expansion step using a domain-specific lexicon, which improved the Noise Sensitivity score from 0.55 to 0.91. Context Utilization was initially low because the LLM would summarize instead of extracting exact figures. Changing the generation prompt to be more extractive improved the score significantly.

Legal: E-Discovery Document Review
A RAG system to help lawyers find relevant information in a large corpus of legal documents.
Objective: High recall and precision in document retrieval.

VI. RAGEVALX APPLICATION

High-Stakes Metrics: Context Recall and Context Precision. It is critical to find all relevant documents (recall) and present the most important ones first (precision).

Findings: A standard retriever configuration had high Context Precision but poor Context Recall. By increasing the number of retrieved documents (top_k) and adding a re-ranking stage, we increased Context Recall by 30% while maintaining high Context Precision for the top 3 results.

VII. DISCUSSION

The case studies highlight the practical utility of the RAGEvalX framework. A single metric is insufficient; a holistic view is necessary for meaningful optimization.

Insights and Practical Considerations

Metric Interdependencies: We observed trade-offs between metrics. For example, increasing top_k to improve Context Recall can decrease Context Relevance by introducing more noise. RAGEvalX helps visualize these trade-offs.

LLM Selection for Evaluation: The choice of the evaluator LLM is critical. For nuanced tasks requiring deep reasoning like Faithfulness and Context Entity Recall, state-of-the-art models like **GPT-4** or **Claude 3 Opus** are recommended. For simpler semantic relevance tasks (Answer Relevancy, Context Relevance), more efficient models like **Llama-3-70B** or **Mistral-Large** provide a good balance of performance and cost.

Integration into MLOps: RAGEvalX is designed for integration into continuous evaluation pipelines. Metrics can be logged using tools like MLflow or Weights & Biases. A subset of robustness tests can be run as part of a CI/CD pipeline to prevent regressions.

Scalability and Deployment
For large-scale evaluation, running LLM-as-a-judge can be costly. We recommend a two-pronged strategy:

Comprehensive Offline Evaluation: Run the full RAGEvalX suite on a golden dataset before deploying a new model.

Lightweight Online Monitoring: In production, track only the Practical Statistics (latency, answer length) and sample a small fraction of traffic for lightweight LLM-based checks like Answer Relevancy.

VIII. CONCLUSION & FUTURE WORK

This paper introduced RAGEvalX, a comprehensive framework for evaluating RAG pipelines across four critical dimensions: Core Accuracy, Context Integrity, Robustness, and Practical Statistics. By providing a structured, multi-faceted approach, RAGEvalX moves beyond simplistic metrics to offer a holistic and actionable assessment of a system's production-readiness. Our case studies demonstrated its value in identifying specific weaknesses and guiding targeted improvements in diverse industrial applications.Future work will focus on three areas. First, we aim to develop more sophisticated, automated methods for generating challenging test cases for robustness evaluation, incorporating techniques from adversarial testing. Second, we will explore the use of smaller, fine-tuned models as evaluators to reduce the cost and latency of the RAGEvalX suite, inspired by the ARES framework. Finally, we plan to extend the framework to evaluate multi-modal RAG systems that incorporate images and tables as context.

REFERENCES

1. P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, 2020.
2. J. Es, et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv preprint arXiv:2309.15217, 2023.
3. J. Saad-Falcon, et al., "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," arXiv preprint arXiv:2311.09476, 2023.
4. S. H. Tu, et al., "Seven Failure Points When Engineering a Retrieval Augmented Generation System," arXiv preprint arXiv:2401.05856, 2024.
5. O. O'Brien, S. Lee, S. Kim, "A Comparison of Data Quality Frameworks: A Review," Data, 2024.
6. C. L. C. Chen, et al., "A framework for extending the health-related quality of life descriptive systems," BMC Medical Research Methodology, 2024.
7. Pradhan, R. and Tomar, G., AN ANALYSIS OF SMART HEALTHCARE MANAGEMENT USING ARTIFICIAL INTELLIGENCE AND INTERNET OF THINGS.
8. Rashmiranjan, Pradhan Dr. "Empirical analysis of agentic ai design patterns in real-world applications." (2025).
9. Pradhan, Rashmiranjan, and Geeta Tomar. "IOT BASED HEALTHCARE MODEL USING ARTIFICIAL INTELLIGENT ALGORITHM FOR PATIENT CARE." NeuroQuantology 20.11 (2022): 8699-8709.
10. Rashmiranjan, Pradhan. "Contextual Transparency: A Framework for Reporting AI, Genai, and Agentic System Deployments across Industries." (2025).
11. Pradhan, Rashmiranjan. “AI Guardian- Security, Observability & Risk in Multi-Agent Systems.” International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1305043.
12. Pradhan, Dr. Rashmiranjan. “Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement.” International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1307013.
13. Wooldridge, M. J. (2009). An introduction to multiagent systems. John Wiley & Sons.
14. Russell, S., & Norvig, P. (2020). Artificial intelligence: a modern approach (4th ed.). Pearson Education.
15. IEEE Std 1012-2016, IEEE Standard for System, Software, and Hardware Verification and Validation. IEEE.
16. IEEE P7000 Standard, Model Process for Addressing Ethical Concerns During System Design. IEEE.
17. Cox, M. T. (2005). Metacognition in autonomous agents. IEEE Intelligent Systems, 20(1), 70-79.
18. Allen, J. F., Ferguson, G., & Stentz, A. (1995). An architecture for integrated planning and reactive execution. In AI planning systems (pp. 1-12).
19. Yao, S., Zhao, W., Yu, Y., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint arXiv:2210.03629. 1
20. Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M. and Amant, R.S., 2021. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Transactions on Artificial Intelligence, 3(6), pp.852-866.
21. Sivakumar, S., 2024. Agentic AI in Predictive AIOps: Enhancing IT Autonomy and Performance. International Journal of Scientific Research and Management (IJSRM), 12(11), pp.1631-1638.
22. Li, J., Qin, R., Guan, S., Xue, X., Zhu, P. and Wang, F.Y., 2024. Digital CEOs in digital enterprises: Automating, augmenting, and parallel in Metaverse/CPSS/TAOs. IEEE/CAA Journal of Automatica Sinica, 11(4), pp.820-823.
23. Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K.S. and Liu, H., 2022. Evaluation methods and measures for causal learning algorithms. IEEE Transactions on Artificial Intelligence, 3(6), pp.924-943.