ISSN: 2769-545X

Research Article

Advances in Machine Learning & Artificial Intelligence

Cost-Efficient Secure Hybrid RAG Assessment

Jitender Singh*

Manager Data Science, Publicis Sapient, India

*Corresponding Author
Jitender Singh, Manager Data Science, Publicis Sapient, India.

Submitted: 2024, Dec 12; **Accepted:** 2025, Jan 27; **Published:** 2025, Jan 31

Citation: Singh, J. (2024). Cost-Efficient Secure Hybrid RAG Assessment. *Adv Mach Lear Art Inte, 6*(1), 01-08.

Abstract

This paper presents a cost-effective and scalable hybrid methodology for evaluating retrieval-augmented generation (RAG) systems using specialized pretrained models and advanced metrics, designed for critical domains like healthcare and finance. LLM judge-based evaluation approaches are hindered by significant score inconsistencies across identical input runs (up to 35% variations) and high computational costs while traditional NLP approaches, overly reliant on entity or phrase matching, lack a multi-faceted perspective and fail to capture deeper semantic understanding. Our methodology addresses these challenges with a novel approach that combines an ensemble of fine-tuned pretrained models and advanced NLP techniques, ensuring consistent, reproducible evaluations across multiple dimensions, while being tailored for offline scenarios to eliminate reliance on internet-connected systems or proprietary LLMs, offering a cost-effective solution with high accuracy in assessing semantic relevance, factual correctness, and context adherence. To enhance reliability, the approach employs a robust weighted scoring system using harmonic means combined with PCA, adaptive, and entropy-weighting techniques for trustworthy and consistent evaluation enabling seamless integration with existing systems, continuous metric adaptation and domain-specific customization, ideal for high-stakes applications that demand rigorous quality assessments without relying on external APIs.

Keywords: Semantic Relevance, Factual Accuracy, Context Consistency, Hallucination Detection, Responsible AI Evaluations, Conversational Systems

1. Introduction

In the rapidly evolving field of retrieval-augmented generation (RAG) systems, evaluating the generator component using online large language models (LLMs) presents significant challenges [1]. LLMs are prone to hallucinations, where they generate information that is factually incorrect or irrelevant, leading to unreliable evaluations. Additionally, these models can introduce scoring biases, affecting the accuracy and fairness of the assessment [2]. This reliance on online models results in non-reproducible outcomes, as the same inputs may yield different results depending on external factors such as updates or model changes [3]. Furthermore, online assessments are often cost-inefficient, requiring continuous API calls and computational resources for large-scale evaluations, making them impractical for many organizations [4].

Traditional NLP evaluation mechanisms, such as precision, recall, BLEU, METEOR, and ROUGE, rely heavily on surface-level entity or phrase matching to assess the quality of generated text [5]. While these metrics are widely used, they fail to capture deeper semantic understanding, which is crucial for tasks like contextual relevance, factual accuracy, and coherence [6]. Precision and recall, for instance, focus primarily on how well entities or keywords match between the predicted and reference text, but they do not account for the meaning or relationships between those entities [7]. This can lead to misleading evaluations, especially in cases where the generated content is semantically correct but does not match the reference text exactly [8].

Metrics like BLEU and ROUGE, which measure n-gram overlap, are also limited in their ability to assess the fluency and coherence of generated text, particularly in more complex or abstract tasks [9]. They might reward models for generating common phrases or sequences without evaluating the contextual appropriateness or factual accuracy of the content. METEOR attempts to improve upon these by considering synonyms and stemming, but it still struggles to evaluate deeper levels of semantic meaning and the overall quality of the response [10]. Consequently, these traditional methods do not provide a holistic view of a model's performance, especially in real-world applications where nuance and accuracy are critical.

Referenced, advanced semantic metrics are crucial for a holistic evaluation of generated text in RAG systems. *Query relevance* ensures the response directly addresses the user's intent, improving the alignment of generated content with the actual query [11]. *Factual accuracy* is important to verify that the information generated is correct, especially in high-stakes domains like healthcare and finance [12]. *Context consistency* ensures the generated content aligns with the broader discourse, preventing contradictions [13]. Semantic *Coherence* measures the logical flow of the output, ensuring readability and comprehensibility [14]. *Semantic relevance* captures the deeper meaning of text, moving beyond surface-level word matching, while *hallucination detection* identifies and mitigates the generation of inaccurate or fabricated information [15]. Together, these techniques provide a comprehensive assessment of both relevance and quality in generated text, offering a more nuanced view compared to traditional metrics.

Offline assessment solutions are critical in healthcare and finance due to security, compliance, and operational concerns [16]. They ensure data privacy by keeping sensitive information within controlled environments, addressing regulatory requirements like HIPAA and GDPR. Offline solutions also mitigate data sovereignty issues by processing data within local jurisdictions [17]. Additionally, they offer cost-efficiency and scalability, avoiding the high costs and limitations of online services. With better reliability and reproducibility, offline evaluations provide consistent, secure assessments, essential for high-stakes decisions in both sectors, such as patient care in healthcare and risk management in finance Rani [18].

Our approach employs established metrics such as relevance, coherence, contextual consistency, factual accuracy, and hallucination detection, but refines them with advanced techniques, thereby enhancing their precision and effectiveness for accurate target metric assessment [19]. For instance, our semantic similarity computation integrates multi-dimensional embedding analysis, cosine similarity, dot product similarity, and context-aware TF-IDF weights [20]. By applying a weighted scoring method, this framework delivers a robust and precise measure of semantic relevance, surpassing the performance of traditional single-metric techniques [21]. Additionally, each metric is computed by evaluating 2–3 sub-metrics, which are then aggregated into a single score. This aggregation process, inspired by ensemble machine learning techniques such as voting, ensures a holistic and accurate assessment aligned with the target intent [22].

2. Methods

2.1 Query Relevance

The query relevance assessment method integrates three distinct techniques to provide a comprehensive evaluation score by capturing semantic, factual, structural, and probabilistic aspects of the query-response relationship and is designed to be adaptable to various query types and response lengths:

- Semantic similarity analysis utilizes sentence embedding models to compute sentence-level semantic alignment, which captures detailed meaning and context correspondence between query and response.
- Knowledge graph analysis identifies concepts by extracting key entities from both query and response using NLP techniques, then relationship mapping is done by constructing graph representations depicting connections between identified entities, where graph comparison analyses similarities between query and response graphs based on: node overlap (shared entities), edge overlap (shared relationships), graph centrality, and additional contextual information in the response.
- Facts consistency analysis extracts and matches similarities between numerical, location, and temporal facts between query, context, and response.

2.2 Factual Accuracy

The factual accuracy assessment method combines four techniques to evaluate response correctness through entity and numerical extraction matched against the provided context:

- Query-response semantic alignment as described in 2.1
- Context-response semantic alignment using equivalent methodology as 2.1, applied to context-response pairs.
- Knowledge graph analysis following methodology in 2.1, applied to context-response entity relationships.
- Facts consistency analysis as detailed in 2.1

2.3 Context Consistency

The context consistency measures how well the response aligns with the context by integrating three distinct techniques to provide a comprehensive evaluation score:

- Coherence analysis employs natural language inference to assess context-response relationships through probabilistic computation of entailment, neutral, and contradiction states, which quantifies the logical consistency between the input context and generated response.
- Context-response semantic alignment using equivalent methodology as 2.1, applied to context-response pairs.
- Knowledge graph analysis following methodology in 2.1, applied to context-response entity relationships.

2.4 Semantic Relevance

The semantic relevance measures overall coverage, contextual coherence, and contextual relevance by integrating four distinct techniques to provide a comprehensive evaluation score:

- Coverage score computes aspect embeddings and segment embeddings by identifying the main topic and aspects from the question and segments/tokens from the answer. Alignment cosine scores are computed using both embeddings and coverage by taking their mean difference from one.
- Coherence score computes taking the mean of cosine similarities of previous steps identified segment embeddings.
- Utilization score uses context embeddings to compute cosines mean against segment and contextual embeddings.

- Relevance score computed by non-linear modeling sigmoid over contextual embeddings.

2.5 Semantic Coherence

The semantic coherence measures the overall coherence and relevance of the response to the given query and context by integrating four distinct techniques to provide a comprehensive evaluation score: NLI coherence and linguistic acceptability (CoLA):

- Computes the coherence between a query-response sentences and context-response sentences pair using a Natural Language Inference (NLI) model, which tokenizes the inputs, passes them through the model, and calculates probabilities for entailment and neutral classifications. The final coherence score is derived from these probabilities, with entailment given full weight and neutral given half weight.
- CoLA method computes linguistic acceptability and coherence scores for query-response and context-response pairs using a pre-trained CoLA model to assess the grammaticality and coherence of the text. The method tokenizes each text, passes it through the model, and obtains probabilities of coherence. It then calculates a weighted average of these scores, with higher weights given to query-response and context-response pairs measuring how well-formed and contextually appropriate the response is.

2.6 Semantic Correctness

The semantic correctness measures the generated response alignment against the query in terms of relevance, factual accuracy, consistency, and response coherence using four techniques to provide a comprehensive score:

- Computes query relevance (ref pt 2.1), factual accuracy (ref pt 2.2), and context consistency (ref pt 2.3)
- Computes the mean pairwise cosine similarity between consecutive segment embeddings, quantifying semantic coherence across sequential text segments.
- Computes a weighted score by integrating query relevance, factual accuracy, context consistency, and response coherence metrics.

2.7 Hallucination Score

The hallucination score assesses the likelihood of hallucinations in generated text, providing a quantitative measure of semantic consistency between the context and the response using two techniques to provide a comprehensive score:

- Natural Language Inference (NLI) uses a pre-trained NLI model to compute a contradiction score between the context and each sentence in the response. Then applies SoftMax to model outputs to obtain probabilities for contradiction, neutral, and entailment classes.
- Question generation and answering technique employs a question generation model to create questions from each sentence in the response, utilizing a question-answering model to generate answers based on the context, then compares the generated answer with the original sentence using semantic similarity.

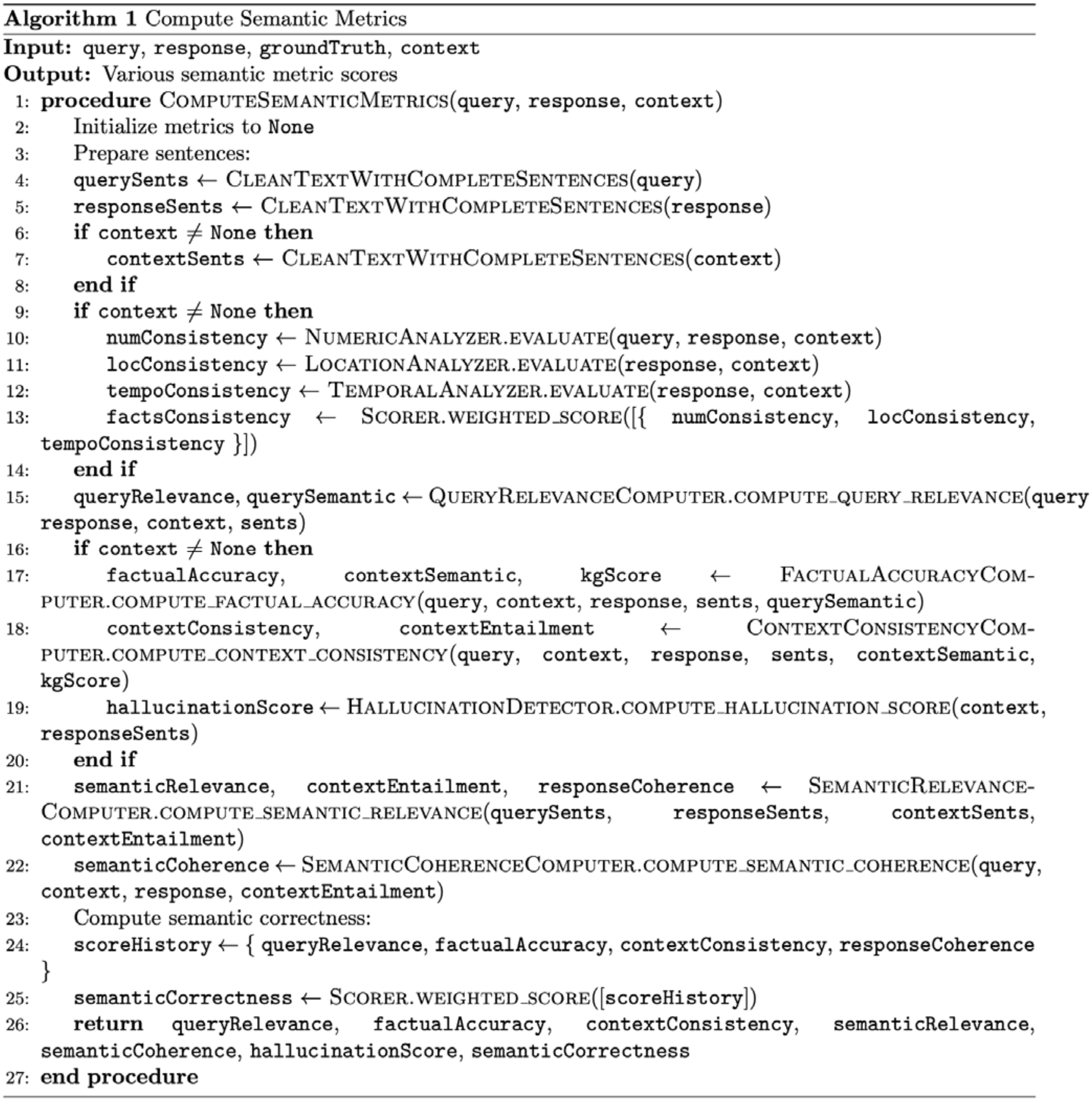
2.8 Generator Competence

The generator competence module computes weighted scores using three strategies:

- PCA (Principal Component Analysis): Analyses variance
- Entropy weighting: Assesses information content
- Adaptive weighting: Evaluates relative variability among scores

The final score combines these weighted approaches with a harmonic mean, penalizing low-performing metrics. This robust mechanism reduces bias from any single metric, providing a comprehensive measure of overall system effectiveness in text generation tasks.

Pseudocode explanation of the algorithm

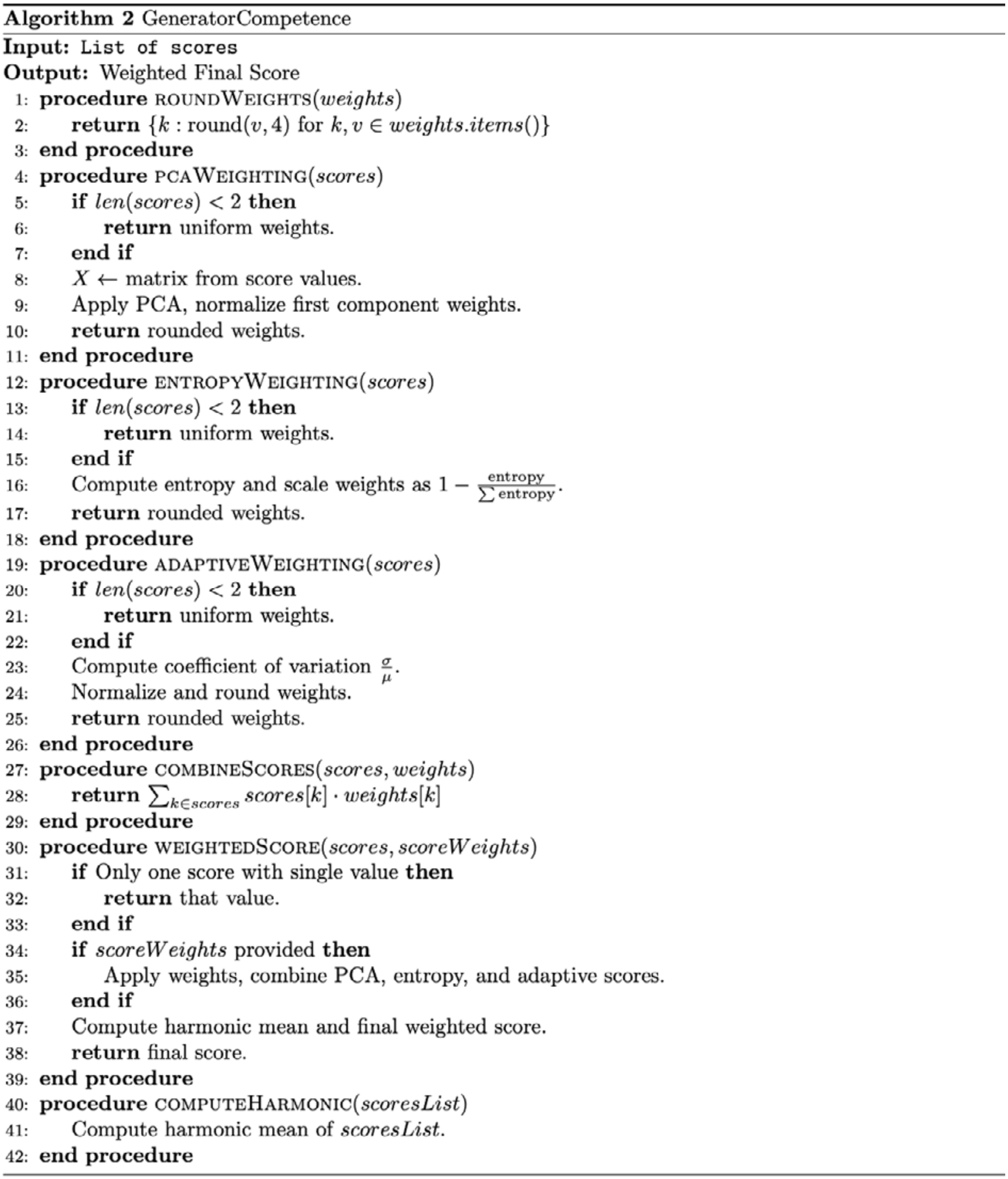


Algorithm 1 Compute Semantic Metrics

Input: query, response, groundTruth, context

Output: Various semantic metric scores

```
procedure ComputeSemanticMetrics(query, response, context)
    Initialize metrics to None
    Prepare sentences:
    querySents $\leftarrow$ CleanTextWithCompleteSentences(query)
    responseSents $\leftarrow$ CleanTextWithCompleteSentences(response)
    if context $\neq$ None then
        contextSents $\leftarrow$ CleanTextWithCompleteSentences(context)
    end if
    if context $\neq$ None then
        numConsistency $\leftarrow$ NumericAnalyzer.evaluate(query, response, context)
        locConsistency $\leftarrow$ LocationAnalyzer.evaluate(response, context)
        tempoConsistency $\leftarrow$ TemporalAnalyzer.evaluate(response, context)
        factsConsistency $\leftarrow$ Scorer.weighted_score([{ numConsistency, locConsistency, tempoConsistency }])
    end if
    queryRelevance, querySemantic $\leftarrow$ QueryRelevanceComputer.compute_query_relevance(query response, context, sents)
    if context $\neq$ None then
        factualAccuracy, contextSemantic, kgScore $\leftarrow$ FactualAccuracyComputer.compute_factual_accuracy(query, context, response, sents, querySemantic)
        contextConsistency, contextEntailment $\leftarrow$ ContextConsistencyComputer.compute_context_consistency(query, context, response, sents, contextSemantic, kgScore)
        hallucinationScore $\leftarrow$ HallucinationDetector.compute_hallucination_score(context, responseSents)
    end if
    semanticRelevance, contextEntailment, responseCoherence $\leftarrow$ SemanticRelevanceComputer.compute_semantic_relevance(querySents, responseSents, contextSents, contextEntailment)
    semanticCoherence $\leftarrow$ SemanticCoherenceComputer.compute_semantic_coherence(query, context, response, contextEntailment)
    Compute semantic correctness:
    scoreHistory $\leftarrow$ { queryRelevance, factualAccuracy, contextConsistency, responseCoherence }
    semanticCorrectness $\leftarrow$ Scorer.weighted_score([scoreHistory])
    return queryRelevance, factualAccuracy, contextConsistency, semanticRelevance, semanticCoherence, hallucinationScore, semanticCorrectness
end procedure
```



Algorithm 2 GeneratorCompetence

Input: `List of scores`

Output: Weighted Final Score

```
procedure roundWeights($weights$)
    return $\{k : \text{round}(v, 4) \text{ for } k, v \in weights.items()\}$
end procedure
procedure pcaWeighting($scores$)
    if $len(scores) < 2$ then
        return uniform weights.
    end if
    $X \leftarrow$ matrix from score values.
    Apply PCA, normalize first component weights.
    return rounded weights.
end procedure
procedure entropyWeighting($scores$)
    if $len(scores) < 2$ then
        return uniform weights.
    end if
    Compute entropy and scale weights as $1 - \frac{\text{entropy}}{\sum \text{entropy}}$.
    return rounded weights.
end procedure
procedure adaptiveWeighting($scores$)
    if $len(scores) < 2$ then
        return uniform weights.
    end if
    Compute coefficient of variation $\frac{\sigma}{\mu}$.
    Normalize and round weights.
    return rounded weights.
end procedure
procedure combineScores($scores, weights$)
    return $\sum_{k \in scores} scores[k] \cdot weights[k]$
end procedure
procedure weightedScore($scores, scoreWeights$)
    if Only one score with single value then
        return that value.
    end if
    if $scoreWeights$ provided then
        Apply weights, combine PCA, entropy, and adaptive scores.
    end if
    Compute harmonic mean and final weighted score.
    return final score.
end procedure
procedure computeHarmonic($scoresList$)
    Compute harmonic mean of $scoresList$.
end procedure
```

3. Architecture

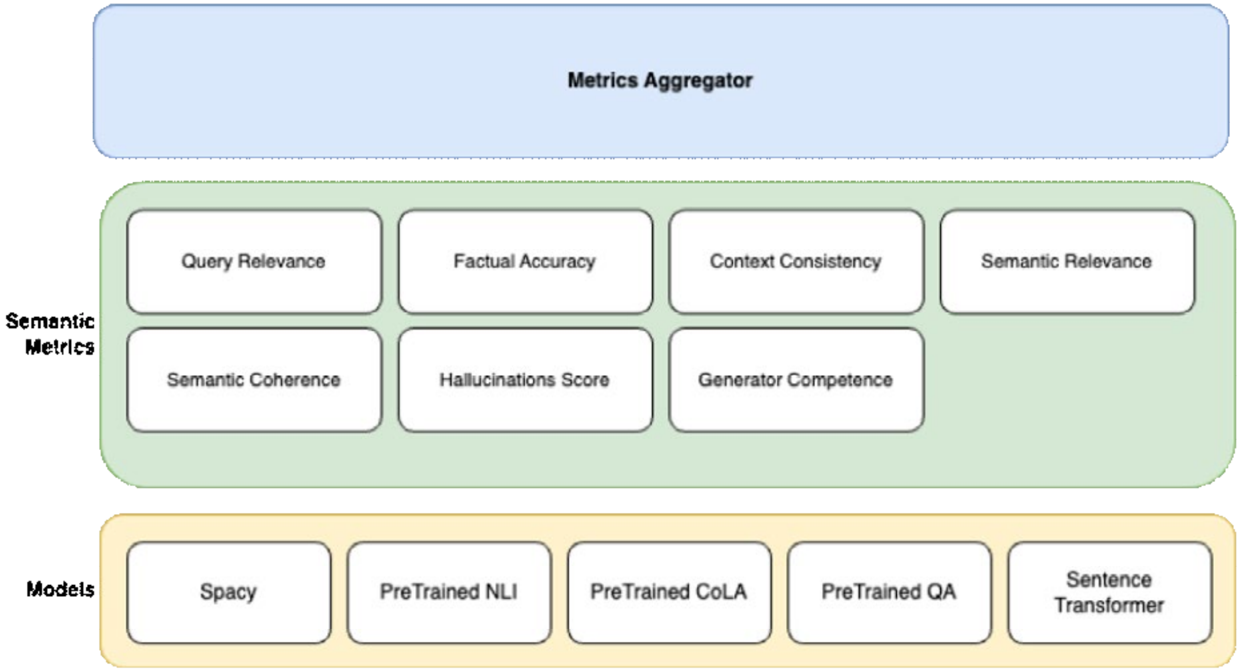


Figure 1: diagram represents a system architecture for evaluating and generating responses using various semantic metrics and language models (LLMs).

Layer by layer explanation of the architectural diagram:

- **Metrics Aggregator Layer:** Collects and combines metrics from various evaluation modules.
- **Semantic Metrics Layer:** Measures various dimensions of the output generated by the model. The key metrics are:

a) Query Relevance: How relevant the response is to the input query.

b) Factual Accuracy: Whether the information provided is factually correct.

c) Context Consistency: Checks if the response is consistent within the given context.

d) Semantic Relevance: Measures how closely the response aligns with the topic.

e) Semantic Coherence: Ensures that the generated text flows logically.

f) Semantic Correctness: Assesses whether the response is appropriate for the asked query.

g) Hallucinations Score: Detects fabricated or incorrect information in the response.

h) Generator Competence: Evaluates how well the text-generation module functions overall.

- **Models Layer:** Contains various models responsible for both understanding and generating responses:

a) SpaCy: Likely used for named entity recognition (NER), part-of-speech tagging, and other language processing tasks.

b) Pretrained NLI (Natural Language Inference): Used to check the logical consistency and infer relationships between the query and the generated response.

c) Pretrained CoLA (Corpus of Linguistic Acceptability): Used to measure linguistic correctness and fluency in the response.

d) Pretrained QA (Question Answering): A pre-trained model for question-answering tasks.

e) Sentence Transformer: Converts input text into dense vector representations for similarity measurement.

4. Results and Discussions

Sr. No	Metric (against Generated Response)	Semantic Evaluation	LLM-as-Judge Evaluation	Percentage Change
1	Query Relevance	0.8825	0.7624	13.61%
2	Factual Accuracy	0.9432	0.8824	6.45%
3	Context Consistency	0.8825	0.7055	20.06%
4	Semantic Coherence	0.8945	0.7704	13.87%
5	Semantic Relevance	0.8025	0.7830	2.43%
6	Semantic Correctness	0.9324	0.8253	11.49%
7	Hallucination Detection	1.0000	0.9601	3.99%
8	Generator Competence	0.9053	0.8127	10.27%

Figure 2: reflects offline metrics under ‘semantic-evaluation’ column against llm-as-judge online approach between reference and generated texts.

The main contributions of this paper are as follows:

- **Cost-Effective, Scalable Offline Evaluation:** Provides a scalable, offline solution tailored for high-stakes sectors like healthcare and finance, reducing external dependencies and ensuring compliance with data privacy regulations, safeguarding and mitigating risks associated with online processing.
- **Enhanced Reliability and Reproducibility:** Ensures stable, reproducible results free from the biases and hallucinations common in online LLM-based evaluations.
- **Multi-Dimensional Assessment:** Utilizes advanced semantic metrics to offer a comprehensive evaluation of generated text.
- **Robust Scoring System:** Leverages harmonic means, PCA, and entropy-weighting techniques to provide a trustworthy, and accurate scoring mechanism.
- **Modular and Customizable:** Allows for continuous adaptation and domain-specific customization, ensuring consistent quality assessments across various NLP tasks.

Conclusion

The proposed hybrid evaluation methodology offers a scalable, cost-effective solution for assessing RAG systems, overcoming challenges such as score inconsistencies and computational inefficiencies. By combining fine-tuned pretrained models, advanced NLP techniques, and a robust weighted scoring system, it ensures accurate, consistent evaluations in critical domains like healthcare and finance. This approach not only supports offline functionality, eliminating reliance on external APIs, but also allows for seamless integration, continuous adaptation, and domain-specific customization, making it ideal for high-stakes applications that require rigorous quality assessments.

References

1. Kau, A. (2024). Understanding retrieval pitfalls: Challenges faced by retrieval augmented generation (RAG) models. *Medium*.
2. Reddy, G.P., Pavan Kumar, Y.V. and Prakash, K.P. (2024). Hallucinations in large language models (llms), *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream),* pp. 1–6.
3. Mehmet Yusuf, E. (2019). Analysis of different data sets of the same clinical trial may yield different results, Annals of *Advanced Biomedical Sciences, 2*(2).
4. Ramírez, G., Lindemann, M., Birch, A., & Titov, I. (2023). Cache & distil: Optimising API calls to large language models. *arXiv preprint arXiv:2310.13561*.
5. Saadany, H., & Orasan, C. (2021). BLEU, METEOR, BERTScore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. *arXiv preprint arXiv:2109.14250.*
6. Ead, W. M., Singh, B. K., Meenakshi, Kumar, A., Srinivas, D., & Sabrol, H. (2024). Enhancing natural language understanding with deep learning: Contextual and semantic implementation. *Linguistic and Philosophical Investigations, 23*(1).
7. Hickling, T. L., & Hanley, W. G. (2005). *Methodologies and metrics for assessing the strength of relationships between entities within semantic graphs* (No. UCRL-TR-216074). Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).
8. Deutsch, D., Dror, R., & Roth, D. (2022). On the limitations of reference-free evaluations of generated text. *arXiv preprint arXiv:2210.12563.*
9. Kim, G., Fukui, K., & Shimodaira, H. (2018, November). Word-like character n-gram embedding. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 148-152).
10. Fay, K., Moritz, C. and Whaley, S. (2023). Considering meaning: What’s it about? what might it really be about?, *Powerful Book Introductions*, pp. 41–67.
11. Bai, L., Guo, J. F., Cao, L., & Cheng, X. Q. (2013). Long tail query recommendation based on query intent. *Chinese Journal of Computers, 36*(3), 636-642.
12. Ni, B., & Huang, Q. Verifying Through Involvement: Exploring How Anthropomorphism Enhances Users’ Intentions to Verify Ai-Generated Information. *Available at SSRN 4963144.*
13. Carnielli, W., & Malinowski, J. (2018). *Contradictions, from consistency to inconsistency* (pp. 1-9). Springer International Publishing.

14. Li, W. (2022) ‘Text genres, readability and readers’ comprehensibility’, European Journal of Computer Science and Information Technology, 10(4), pp. 52–62.
15. Li, X., Gao, Y., Li, W., & Yang, L. (2024, March). A Text Generation Hallucination Detection Frame-work Based on Fact and Semantic Consistency. In 2024 *5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)* (pp. 970-974). IEEE.
16. Gonzalez-Granadillo, G., Diaz, R., Karali, T., Caubet, J., & Garcia-Milà, I. (2021). Cyber-Physical Solutions for Real-time Detection, Analysis and Visualization at Operational Level in Water CIs. *CYBER-PHYSICAL THREAT INTELLIGENCE FOR CRITICAL INFRASTRUCTURES SECURITY*, 188.
17. Tzanou, M. (2020). Addressing big data and AI challenges: A taxonomy and why the GDPR cannot provide a one-size-fits-all solution. In *Health Data Privacy under the GDPR* (pp. 106-132). Routledge.
18. Rani, S., Kataria, A., Bhambri, P., Pareek, P. K., & Puri, V. (2024). Artificial Intelligence in Personalized Health Services for Better Patient Care. In *Revolutionizing Healthcare: AI Integration with IoT for Enhanced Patient Outcomes* (pp. 89-108). Cham: Springer Nature Switzerland.
19. Kaifi, R. (2024). Enhancing brain tumor detection: a novel CNN approach with advanced activation functions for accurate medical imaging analysis. *Frontiers in Oncology, 14*, 1437185.
20. Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2022). Netflix recommendation system based on TF-IDF and cosine similarity algorithms. *no. Bml,* 15-20.
21. Wu, D. (2019). Robust Face Recognition Method Based on Kernel Regularized Relevance Weighted Discriminant Analysis and Deterministic Approach. *Sensing and Imaging, 20*(1), 36.
22. Tran, T., Tsai, P., Jan, T., & Kong, X. (2010). Network intrusion detection using machine learning and voting techniques. *Machine Learning*, 267-290.