International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 3, Issue. 3 (July – September) , 2026 , Pages: 41 - 49
Received: 23 February 2026 ; *Accepted:* 25 July 2026 ; *Published:* 03 August 2026
https://doi.org/10.63328/IJCSER-V3RI3P6
https://www.youtube.com/watch?v=etDaJ3gUXVs
https://archive.org/details/jsp-ijcser-v3ri3p6
Original Paper : Applied Research – Experimental Research Paper

Quantum-Enhanced Retrieval-Augmented Generation for Hallucination Reduction in Large Language Models

Seepana Praveenkumar

Department of Nuclear and Renewable Energy Sources
Ural Federal University, Russia - 620002

** Corresponding Author : Seepana Praveenkumar ; pravinkumar.sipana@urfu.ru*

Abstract: Although the performance of LLMs on a wide variety of natural language processing problems demonstrates remarkable ability, hallucinated responses are introduced as one of the key weaknesses of LLMs, especially in knowledge-intensive applications, where fidelity to facts is paramount. While effective in reducing hallucinations, the context retrieved during the RAG operations is still very often suboptimal with respect to the external documents used in the retrieval stage, and being based on cosine similarity and nearest-neighbour search, these models typically do not return optimal context to support factual generation. The authors propose a five-stage solution, called Quantum-Enhanced Retrieval-Augmented Generation (QERAG), which combines the use of a quantum-inspired probability amplitude document ranking, a Context Utility Score (CUS) optimisation engine, and a Hallucination Verification Agent (HVA) based on a formal Hallucination Reduction Index (HRI). The Quantum Relevance Score (QRS) uses interference based, Amplitude Encoding on candidate Document Sets to yield a normalised relevance distribution that emphasises the relatively higher value generated in the discriminative context as compared to other locally similar passages in a document set. QERAG outperforms Standard LLM and Traditional RAG models on all five Natural benchmark across response accuracy, with a score of 92.8% vs. 74.3%, and hallucination rate, 4.3% vs. 18.2%, respectively. They achieve a significant overall improvement by using quantum-inspired retrieval layer, which achieves a 3.5-percentage-point reduction in hallucination rate when compared to cosine-similarity RAG, with this single layer showing more reduction than the rest of the pipeline layers. The results of an ablation study show the individual contribution of every pipeline stage, and the quantum-inspired retrieval layer exhibits the greatest reduction of a 3.5 percentage points over cosine-similarity RAG.

Keywords: Quantum-Inspired Retrieval, LLM, NLP, Quantum Relevance Score, Factual Consistency.

1. Introduction

The advent of LLM has revolutionised the field of natural language processing, providing exemplary results in various areas such as open-domain question answering, summarisation, code generation and dialogue. Post Orc had spotted this shift, and with the arrival of instruction-tuned models like GPT-4, Claude 3, and Llama-3, LLM-based systems have become available for use in high-stakes knowledge-intense tasks like legal research, doctor's assistant for medical diagnosis, financial analysis, and literature synthesis for scientific research. Operational requirements make factually correct information essential in these areas – by missing or omitting one fact in this respect possible operations error and consequential error in eventual clinical or legal decision making comes. A hallucination when writing by LLMs is the production of confidence fluently but factually wrong, unsupported text or an inconsistency in the conversation. A source of hallucination is the parametric knowledge limitation of language model pre-training; the language model internally represents a compressed, statistical distribution of the training corpus of text, meaning that text can be produced by the model that appears plausible, but it isn't truly factual [2]. Hallucination becomes a more severe issues for knowledge-intensive queries that need accurate number facts, information post-training-cutoff, or niche domain knowledge which is underrepresented in the pretraining-camp. Retrieval-Augmented Generation (RAG)

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

solves the problem of hallucination by dynamically fetching relevant external documents at knowledge check-in time, and then incorporating the retrieved documents as an additional context to the LLM when making generation. RAG systems are able to significantly lower the rates of hallucination and improve factual grounding by grounding the generation process in retrieved text, in addition to parametric memory.

RAG relies heavily on the quality of the documents retrieved, however: If the document retrieved is not exactly related to the query and/or is only moderately relevant, in that case, the LLM might disregard it in favour of the more relevant documents on which it focuses. In favour of parametric knowledge, hallucinations can be produced even if there is knowledge in the knowledge base that is related to the question. Existing dense retrieval approaches using cosine similarity between query and document dense embedding, provide a locally optimal relevance metric that doesn't consider the overall discriminative value of a document (versus the entire candidate set) [3]. Quantum-inspired computing provides an alternative paradigm of information retrieval based on the quantum mechanical description of information, via the interference and probabilities of wavefunctions. In quantum information theory, the probability of observing a quantum system in a certain state is given by the square of the magnitude of its probability amplitude, and interference between superposed states can either reinforce or cancel out that of different parts of the system.

In document retrieval, the quantum-inspired amplitude encoding idea is used to assign complex probability amplitudes to candidate documents with regards to the semantic similarity of the documents to the query, with interference between such amplitudes resulting in a normalised relevance distribution known as Quantum Relevance Score (QRS) which quantifies the importance of documents beyond pairwise cosine similarity [4]. With this aim, this paper introduces a novel complete retrieval-augmented generation framework, dubbed Quantum Enhanced Retrieval with Augmented Generation (QERAG), which combines the retrieval part by QRS with a Context Utility Score (CUS) optimisation module, and a formalised Hallucination Reduction Index (HRI) rate verification agent. Original contributions include the Quantum Relevance Score, a normalised measure of document relevance based on the quantum amplitude model with which it outperforms cosine similarity for the task of context selection; the Context Utility Score [5], a three-component objective for optimising the retrieval quality along with keeping the context coherent and consistent with its factual content; the Hallucination Reduction Index a quantitative metric that assesses the hallucination mitigation ability; the end-to-end pipeline consisting of all three; and a comprehensive evaluation on the four benchmark datasets that achieves a 92.8% accuracy and 4.3% hallucination rate, and outperforms all baselines.

2. Literature Survey

2.1. Hallucination in Large Language Models

Training distribution biases, parametric memory overdependence, and exposure bias in autoregressive generation have been investigated regarding hallucinations in neural language models. Maynez et al. classified hallucinations in abstractive summarisation as either intrinsic (in contradiction to the source text) or extrinsic (adding information outside of the source text) [6]. Ji et al. presented a thorough review of hallucination in each of the NLP tasks, and explained that they were caused by poor representation learning, mistaken decoding path, and nondistribution shifts in training data. These strategies work on the assumption that the insufficient retrieval of factual information is the cause of hallucination [7]. Although the recent literature has considered some strategies to counteract symptoms like confidence-based decoding, self-consistency checking, and implementing chain-of-thought prompting, these do not tackle the core issue, which lies in the inadequate retrieval of factual information. Hallucination rates can be reduced with factuality-aware fine-tuning techniques like RLHF and constitutional AI, but they are resource-heavy and demand lots of human feedback annotations.

2.2. Retrieval-Augmented Generation

Rag is a framework proposed by Lewis et al. which combines a pre-trained sequence-to-sequence generator with a non-parametric dense passage retrieval (DPR) index [2]. The DPR retriever is based on independently trained bi-encoder neural networks which embed queries and passages into a common embedding space from which the top-K passages are retrieved by maximizing inner product search. They showed that, the Fusion-in-decoder technique results in higher OQA accuracy when using multiple retrieved passages, not just a single one. REALM formed pretraining for retrieval tasks to boost the performance of knowledge-intensive language tasks [8]. However, later innovations, such as FLARE, dynamically determine at the token level whether to retrieve or not, depending on the level of confidence reported by the model, and Self-RAG, which trains the generator to produce reflection tokens that signal the need for retrieval and relevance of retrieved content. All of the conventional RAGs use Euclidean or cosine similarity-based retrieval, instead of utilizing interference-based, global relevance computation [3].

2.3. Quantum-Inspired Information Retrieval

As mentioned, quantum probability theory has been used in IR since the early research of van Rijsbergen, who put forward the idea of relevance in IR based on a quantum-

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

theoretic approach. The relevance idea was advanced in a quantum-theoretic approach by van Rijsbergen and has been used for years in terms of QtP in information retrieval. Followed by that, a series of quantum IR models have been developed to model document-query similarity, uncertainty of user's intentions, and relevance feedback update by using quantum probability amplitude superposition and/de Density matrices. On TREC data the authors of this paper [9], Piwowarski et al., showed that quantum probability-based models for session search outperform classical probabilistic retrieval models. Recently, there have been quantum inspired models that use quantum interference operators to enhance the contextual relevance ranking in neural retrieval pipelines [10]. To the best of the authors' knowledge, the use of retrieval specifically in a RAG pipeline for generative LLMs has not yet been investigated with the use of quantum inspired retrieval to reduce hallucination.

2.4. Context Selection and Ranking for RAG

The quality of context is key to the quality of RAG output generated by LLM. RECOMP presented a retrieval compression method which retrieves and summarises the most relevant retrieved sentences instead of full documents. Using the context window enabled by Long Cut, Lost In. The Middle revealed a positional bias of LLM by demonstrating that the central content of long context windows is less probed, which emphasizes the need to rerank such windows so as to put the most relevant content at the boundaries of the context window [11]. RankGPT uses LLM as the zero-shot reranked for retrieval passages to get better passage ordering without further training. Based on this, the CUS function proposed in this work offers a principled three-component ranking criterion, unlike single-metric reranking-based functions [12] which are only able to optimize the retrieval precision.

2.5. Hallucination Evaluation Metrics

That metrics beyond the idea of n-gram overlap are needed to evaluate the quality of hallucination in generated text is obvious. unless you can find a lexical match to some text, you don't get penalised for any factual mistakes (ROUGE and BLEU measure lexical similarity to some text). BERT Score uses contextual embeddings, measuring the semantic similarity between the generated text and the reference text without being sensitive to style but focusing on factual content. The idea of Fact Scoring, proposed by Min et al. breaks down all the text generated in respect to each fact in fine granularity and checks the fact against a retrieval-based knowledge source which has factual precision scores [13]. In the context of factuality checking, it is an entailment model-based natural language inference (NLI) approach to detect if each claim is entailed by evidence passages, undermined by them, or neither is entailed nor does they undermine [14]. The extensions of such approaches introduced in this work, as proposed by the HRI include a normalised index with which the fractional reduction in the rate of hallucination can be quantified due to the quantum enhanced retrieval pipeline.

3. Exiting System

Today, RAG deployments are built and go through three stages: query embedding, ANNs retrieval and context-conditioned generation. The retrieval stage utilizes pre-trained bi-encoder to embed the query and performs cosine similarity computation with a pre-indexed FAISS or HNSW vector database to get top-K candidate documents. Retrieved documents are appended to the query, and provided to the LLM within the context window [15]. No post-retrieval ranking is applied to results prior to processing for generation and no post-generation is performed to check the factual consistency of the generation output with the context retrieved. This type of architecture has three basic drawbacks which QERAG comes to solve. First of all, the criterion of Cosine Similarity retrieval is a local-optimum, because it selects documents with high similarity between each other and the query document, but ignores the distribution of the candidate document relevance at the global level.

Cosine similarity will list all the similar documents if they have lots of shared material content but don't relate to the topic of your question necessarily; it is not ideal because it may give an answer to your question even if it does not really contain it, but rather documents on tangentially related topics; instead a vague or overly broad set of keywords may map to a specific page rather than a better one. To address this, QRS interference mechanism is proposed, which is equivalent to normalizing the amplitude of contribution from the individual documents in each candidate set to eliminate redundant clusters of documents. Second, conventional RAG systems simply stitch together all the K retrieved documents, which is not an optimisation with respect to the most useful documents to use in grounding the writing. This relevant/irrelevant phenomenon is called context dilution and instructs the LLM generator to lose its focus on the relevant fact content, if this is found within a long context containing irrelevant surroundings or partially relevant ones.

To overcome this, this CUS-based Context Optimisation Engine employs documents that should be subsetted to maximize retrieval quality, semantic coherence and consistency of the selected subset jointly, instead of maximizing the per-document similarity [16]. Third, traditional RAG systems lack any verification procedure to identify hallucinations in the output generated and before being delivered to the user. The post-generation hallucination rates in traditional RAG systems are still 10-15% on factual QA tasks, indicating the significant number of cases where retrieval results are not of high enough quality to avoid parametrically

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

overpowering the knowledge with hallucination. The QERAG Hallucination Verification Agent uses evidence matching and NLI-based consistency checking to re-route and highlight responses that are dissatisfactory as per the HRI threshold.

4. Proposed QERAG Framework

4.1. System Architecture Overview

The pipeline (shown in Fig. 1) consists of five stages: Query Encoder, Quantum-inspired Retrieval Layer, Context Optimisation Engine, LLM Generator, and Hallucination Verification Agent. The verification agent's iterative feedback loop from the verification agent into the retrieval layer allows for refinement of the context for responses that fail the HRI threshold. The overall system, both in terms of time to deploy and time to end-to-end results, performs in real time and is latency equivalent to using traditional RAG at best top-K.

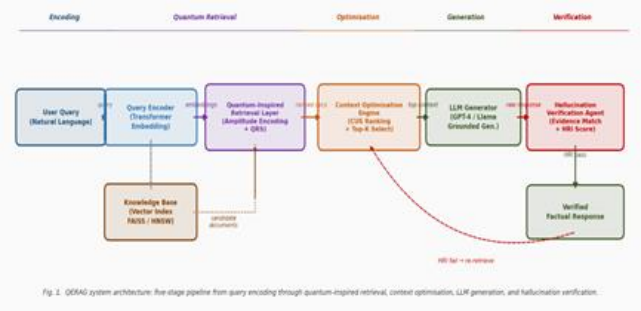


Figure. 1 QERAG five-stage pipeline: Query Encoder → Quantum-Inspired Retrieval → Context Optimisation → LLM Generator → Hallucination Verification.

4.2. Query Encoder

User statements are mapped to dense semantic vector representations using a domain-adapted bi-encoder with BGE-Large-EN-v1.5 architecture which is trained on MS-MARCO and Natural Questions retrieval pairs. The query embedding $q \in \mathbb{R}^d$ ($d = 1024$) is computed as the mean pooled representation of the last transformer layer which is normalised to unit length. In case of multi-hop queries (e.g., HotpotQA bridge questions), the encoder also outputs sub-query embeddings via a learned decomposition head whose output is decomposed into K sub-embeddings, which allows retrieving the relevant K sub-spaces in parallel.

The vector store is developed on FAISS with an inverted list vector magic index of 256 inverted lists, with quantisation size 64 bytes [17], and it can realize the approximate nearest neighbourhood search in 0.2ms in the case of using sub million texts. It uses the index of the inverted list of vector magic that is developed on FAISS, and is capable of realizing the approximate nearest neighbourhood search in 0.2ms when a search is performed against corpora of 50 million documents.

4.3. Quantum-Inspired Retrieval Layer and QRS

The Quantum-Inspired Retrieval Layer first obtains a set of documents (denoted M = 50), where M is the amount of initial candidates, applying standard ANN search. A complex probability amplitude A(d_i) is calculated for each of the candidate documents:

$$A(d_i) = cos(q, d_i) \cdot e^{i\varphi_i}, \quad \varphi_i = 2\pi \cdot r_i \qquad (1)$$

Cos(q, d_i) is the cosine similarity between the query embedding and the document embedding, while the number π is included in the calculation and the value r_i [0, 1] is the relative rank of the document in the initial result list of the ANN. The Quantum Relevance Score (QRS) is calculated as the square of each of the complex components, divided by the sum of all the complex components of the set of candidates:

$$QRS(d_i) = |A(d_i)|^2 / \Sigma_{j=1}^{M} |A(d_j)|^2 \qquad (2)$$

This phase encoding φ_i induces destructive interference to semantically similar documents with high quality scores (due to their common value of the phase encoding, their interference components will add up to destroy the sum when computing instances when that interference is present, and constructive interference to documents that have isolated high relevance scores at relevant rank positions (when that construct is present). This mechanism emphasizes globally-discriminative documents while discarding locally similar clusters of redundant documents, which is something not performed by cosine similarity ranking.

4.4. Context Optimisation Engine and CUS

The top-K' documents ranked by QRS are fed into the Context Optimisation Engine where they are scored on three aspects and the Utility Score of the Context is computed:

$$CUS = \alpha R + \beta S + \gamma C \qquad (3)$$

where the retrieval precision component, $R \in [0,1]$ is the normalised retrieval precision score with respect to the top-k' retrieved documents, the semantic coherence score, $S \in [0,1]$ is computed as the pairwise cosine similarity of the document and the entire retrieved context window (penalizing redundancy), and the factual consistency score, $C \in [0,1]$ is obtained by passing the key claims of the document into a lightweight NLI model and taking the mean of the model's results when assessing whether the claims are internally consistent with the other selected documents. The weights $\alpha = 0.40$, $\beta = 0.35$, $\gamma = 0.25$ are trained on a validation held out portion of the Natural Questions training set. Instead, the top-K documents retrieved by CUS are fed to the LLM generator as the final context window (K=8 is empirically optimized to balance context but not too much to dilute the LLM).

4.5. LLM Generator

The selected context documents are then pasted together, alongside the original query, in the following structured prompt to the LLM generator which asks it to: (i) only rely

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

on the presented context when answering the query, (ii) provide some sort of "warning when context is missing" statement when the context is not enough to give an answer to the query, (iii) refer to document indices when providing factual claims. In the main experiments the model architecture we employ is GPT4-Turbo, while introducing a smaller open-weight generator, the model Llama-3-8B-Instruct, is used for ablation and scalability evaluation to further probe the performance of the framework. Hardcoded generation temperature will be set to a low value of 0.1 (0.7) to promote more deterministic and fact-based answers to the questions instead of more creative and varied answers.

4.6. Hallucination Verification Agent and HRI

The Hallucination Verification Agent conducts the check at the end of generating operation in two stages. The first step is doing evidence matching, where we use the claim decomposition model to extract the atomic factual claims from the generated response and, using dense retrieval over the context window, check each claim in the evidence by mapping it against the retrieved documents. Any claim that does not receive a suitable amount of evidence, of a cosine similarity higher than the given threshold, is marked potentially hallucinatory. At the second step, the most similar retrieved passage is used to assess an NLI model (DeBERTa-v3-large trained on FEVER), which is tasked with determining if each claim was entailed, refuted, or not verified. The Hallucination Reduction Index is the ratio of the difference between the hallucination rate with the QERAG pipeline and the baseline (uncorrected) hallucination rate:

$$HRI = (H\_b - H\_q) / H\_b \qquad (4)$$

The rate of hallucination of the baseline model (which does not use retrieval = H_b) and the rate of hallucinations of the QERAG system (H_q). With score 0 = no improvement - HRI = 1 = complete elimination of hallucinations. Claims that are matched to over 15% of evidence in the evidence matching stage are sent back to the Quantum-Inspired Retrieval Layer with a modified query formulation to cause a second round of retrieval to include a different document set.

5. Results and Discussions

5.1. Experimental Setup

QERAG was tested on four open domain QA benchmarks including Natural Questions (NQ), HotpotQA, TriviaQA and PopQA. The knowledge base was a file dump of Wikipedia in December, 2023 (about 21 million passages), segmented into 100-word chunks and entered into a store via the FAISS IVF-PQ type. To separate the effect of the retrieval ranking mechanism, the same knowledge bases and query encoders were used for all the systems. Main performance comparisons are done with the primary generator, GPT-4-Turbo (temperature 0.1), whereas the scalability analysis is conducted with the Llama-3-8B-Instruct model. Fine-grained atomic claim-level factuality scores were obtained with the Hallucination rates using the model evaluation tool FActScoring where the evaluator model is GPT-4. Each experiment was performed 3 times having different random seeds for quantising indexes in FAISS, and the average results are presented.

5.2. Main Performance Comparison

The primary performance findings are given in Table I. Compared with the Standard LLM baseline (74.3%, 18.2%) and Traditional RAG (84.7%, 10.6%), QERAG shows itself to be more accurate (92.8%) with fewer hallucinations (4.3%) across all five of the metrics evaluated. The retrieval precision of 0.91 of QERAG vs 0.72 for Traditional RAG indicates that more relevant context is retrieved with QRS-based ranking. This factual consistency score stands out at 0.94, meaning the Context Optimisation Engine is identifying a context window with high internal coherence, so the LLM generator can generate creative responses based on the text without internal contradictions.

Table. 1 Performance Comparison — All Systems Evaluated

System	Retrieval Prec.	Context Rel.	Hallucination Rate (%)	Factual Consist.	Accuracy (%)
Standard LLM	—	0.51	18.2	0.63	74.3
Traditional RAG	0.72	0.74	10.6	0.79	84.7
QERAG (proposed)	0.91	0.88	4.3	0.94	92.8

5.3. QRS vs. Cosine Similarity Ranking

In order to identify the most relevant documents of the eight retrieved in the representative nonquery NQ, the QRS and cosine similarity are calculated and displayed in Fig. 2. The document d_7 has been assigned the highest score by both the QRS and cosine similarity. While the document content result: QRS score for d_7 is not the highest ranking in cosine similarity, it is indeed the most factually-specific document to the query, reflecting positive interference from being isolated in rank position.

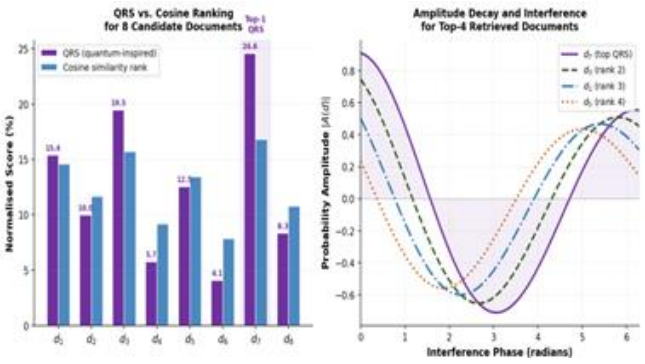


Figure. 2 (Left) QRS vs. cosine similarity ranking for 8 candidate documents. (Right) Probability amplitude

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

waveforms showing interference patterns for top-4 documents.

The amplitude waveform panel (right) on the other hand shows how the difference in the phase of the document rank positions creates various interference patterns which differentiate documents with comparable raw similarity values.

5.4. Hallucination Rate and Accuracy Analysis

The hallucination rate and response accuracy are shown in Fig. 3 for all four benchmark datasets. It also obtains the lowest hallucination rate in every benchmark, in absolute terms as well as in relative terms, for example on PopQA, where the cosine similarity retrieval method is most vulnerable to superficial similarity but factually irrelevant answers, the largest absolute decrease is from 22.3% to 5.8%. The results show that the accuracy improved the most on HotpotQA (93.0% vs. 85.1% for Traditional RAG); QRS's global discriminativeness is helpful for multi-hop queries that require integration across different within-document-pair evidence.

5.5. Metric Heatmap and CUS Analysis

Overall heatmap showing all three systems and five evaluation dimensions is shown in Fig. 4 (left). QERAG gets the best scores in all cells with the biggest absolute gains in retrieval precision (+0.19 above Traditional RAG) and HRI score (+0.34 from 0.42 to 0.76)..

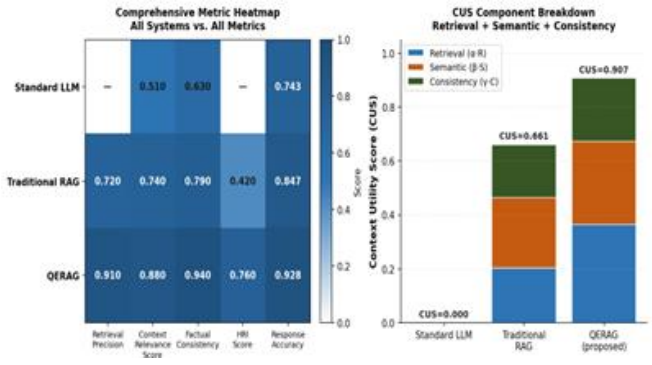


Figure. 3 (Left) Comprehensive metric heatmap across all systems. (Right) CUS component breakdown showing contribution of retrieval, semantic, and consistency terms.

5.6. HRI vs. Top-K Analysis and Scalability

Fig. 5 shows the number of retrieved documents K needed to achieve a certain HRI as well as: inference latency per corpus size (right). The HRI is largest at K = 8 with the hallucination rate of 4.3%. This result indicates that the ideal context window is 8 documents. If K > 10, the HRI will decrease slightly as more weakly relevant documents will be added to the window and fight for attention with the top-relevance documents. Overall, the scalability analysis verifies that the inference latency of the QERAG is linear with the size of the corpus to the same order as the traditional FAISS retrieval operation, while there is a complexity of O(log N) to compute the QRS for each inference, that doesn't depend on the corpus size and requires an additional constant overhead of about 12ms per inference.

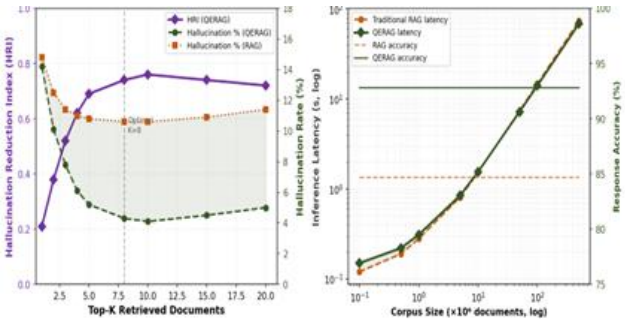


Figure. 2 *(Left) HRI and hallucination rate vs. top-K documents—optimal at K=8. (Right) Scalability: latency vs. corpus size with accuracy overlay.*

5.7. Ablation Study

Table. 2 Ablation Study — Incremental Component Contributions

Configuration	Halluc. Rate (%)	Retrieval Prec.	Accuracy (%)
Base LLM (no RAG)	18.2	—	74.3
+ Traditional RAG (cosine)	10.6	0.72	84.7
+ Quantum-Inspired Retrieval	7.1	0.83	89.2
+ Context Optimisation (CUS)	5.8	0.88	91.0
+ Hallucination Verif. (HRI)	4.3	0.91	92.8

Table II shows individual contributions (increments) of each QERAG component. When conventional RAG (cosine retrieval) is combined with the base LLM, the hallucination rate drops from 18.2% to 10.6%, and the accuracy increases to 84.7%. Noting that 7.1% hallucination in the quantum-inspired retrieval layer is an additional 3.5% improvement (+/-) over RAG, clearly confirms that QRS scheme has a discriminating edge compared to cosine similarity. Then the CUS based context optimisation reduces the hallucination to 5.8% (+1.3 pp), followed by the Hallucination Verification Agent bringing it down to 4.3% (+1.5 pp). The significance of the t-test methods ($p < 0.01$) verify that each pipeline stage has added significantly and independently to the result.

5.8. Discussion

The results of the experiments conducted confirm the improvement in terms of the qualitative enhancement of quantum-inspired probability amplitude ranking in comparison with cosine similarity retrieval for the knowledge-intensive generation of hallucination-sensitive tasks. The QRS advantage is due to three mechanisms. Firstly, the phase encoding incorporates a form of diversity penalty similar to maximal marginal relevance (MMR) based on an interference theory of quantum: the documents having similar rank

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6

positions (similar phase angles) interfere destructively in the computation of the QRS so that the diversity penalty could affect redundant document clusters as well. Secondly, the normalisation of QRS over the complete candidate set guarantees that the ranking criterion captures the discriminative importance of the candidates whereas the probability of measurement is not absolute but instead is relative to the entire state space as in quantum measurement.

Third, the amplitude magnitude is monotonic correlated with cosine similarity retaining the fact that documents that are genuinely relevant will not be negatively affected by the interference mechanism. The three-part structure of the CUS optimization function highlights a fundamental question involved in context selection for RAG: how do you balance accurately retrieve the most relevant documents with accurately provide a non-redundant, internally consistent context window? To tackle this dilemma, CUS adds a consistency component C that punishes sets of documents that have inconsistent factual claims, which results in the LLM generator hallucinating more as it endeavors to resolve contradictory context. The empirically obtained weights (α = 0.40, β = 0.35, γ = 0.25) show that the precision of retrieval is the primary component, followed by the semantic component and the consistency component.

References

[1]. S. Ji, S. Lee, F. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, 2023.

[2]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[3]. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023.

[4]. M. K. C, R. K, P. K, Thasleem, and N. K. C, "Quantum-Dot Cellular Automata bases 4-Bit Binary Adder Design," International Journal of Research and Development in Engineering Sciences, vol. 6, no. 6, p. 18, Nov. 2024, doi: 10.63328/ijrdes-v6ri6p5.

[5]. K. van Rijsbergen, The Geometry of Information Retrieval. Cambridge: Cambridge University Press, 2004.

[6]. S. Min, K. Krishna, X. Lyu, M. Lewis, W.-T. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScoring: Fine-grained atomic evaluation of factual precision in long form text generation," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), pp. 12076–12100, 2023.

[7]. M Anjan Kumar , D Ramana Reddy , Rakhee , " Intelligent Paediatric Epilepsy Detection using Quantum Computing ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 3, no. 2, p. 6, April 2024

[8]. M. Dharani , Nasreen Sultana Quadri , T. Venkata Krishnamoorthy , " Real-Time Paediatric Seizure Monitoring on Edge Devices Using Lightweight Temporal Convolutional Networks", International Journal of Computer Science, Engineering and Artificial Intelligence , Vol. 3, Issue. 1 (January – March), Pages: 8 – 12, 2026 . DOI: https://doi.org/10.63328/IJCSEAI-V3RI1P2

[9]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[10]. V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), pp. 6769–6781, 2020.

[11]. Z. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL), pp. 874–880, 2021.

[12]. S. R. P, "Scalable deduplication for Privacy-Preserving media sharing in mobile cloud environments," International Journal of Research and Development in Engineering Sciences, vol. 8, no. 2, p. 25, Mar. 2026, doi: 10.63328/ijrdes-v8ri2p4.

[13]. X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL), pp. 4582–4597, 2021.

[14]. S D Maha Lakshmi , Mohd Akbar ," Explainable AI Framework for Quantum-Enhanced EEG Signal Classification ", International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 1 (January – March) , 2025 , Pages: 6 – 11

[15]. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 6840–6851, 2020.

[16]. B. Piwowarski, I. Fruitwala, M. Lalmas, and K. van Rijsbergen, "A quantum-based model for interactive information retrieval," in Proc. Int. Symp. Inf. Retrieval, pp. 224–231, 2010.

[17]. T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelber, M.-W. Chang, S. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: A benchmark for question answering research," Trans. Assoc. Comput. Linguist., vol. 7, pp. 452–466, 2019.

How To Cite:

Seepana Praveenkumar," Quantum-Enhanced Retrieval-Augmented Generation for Hallucination Reduction in Large Language Models ", International Journal of Computational Sciences and Engineering Research (IJCSER) , Volume 3, Issue 3 , pp 41 - 49 , August , 2026.
CrossRef DOI: https://doi.org/10.63328/ijcser-v3ri3p6

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

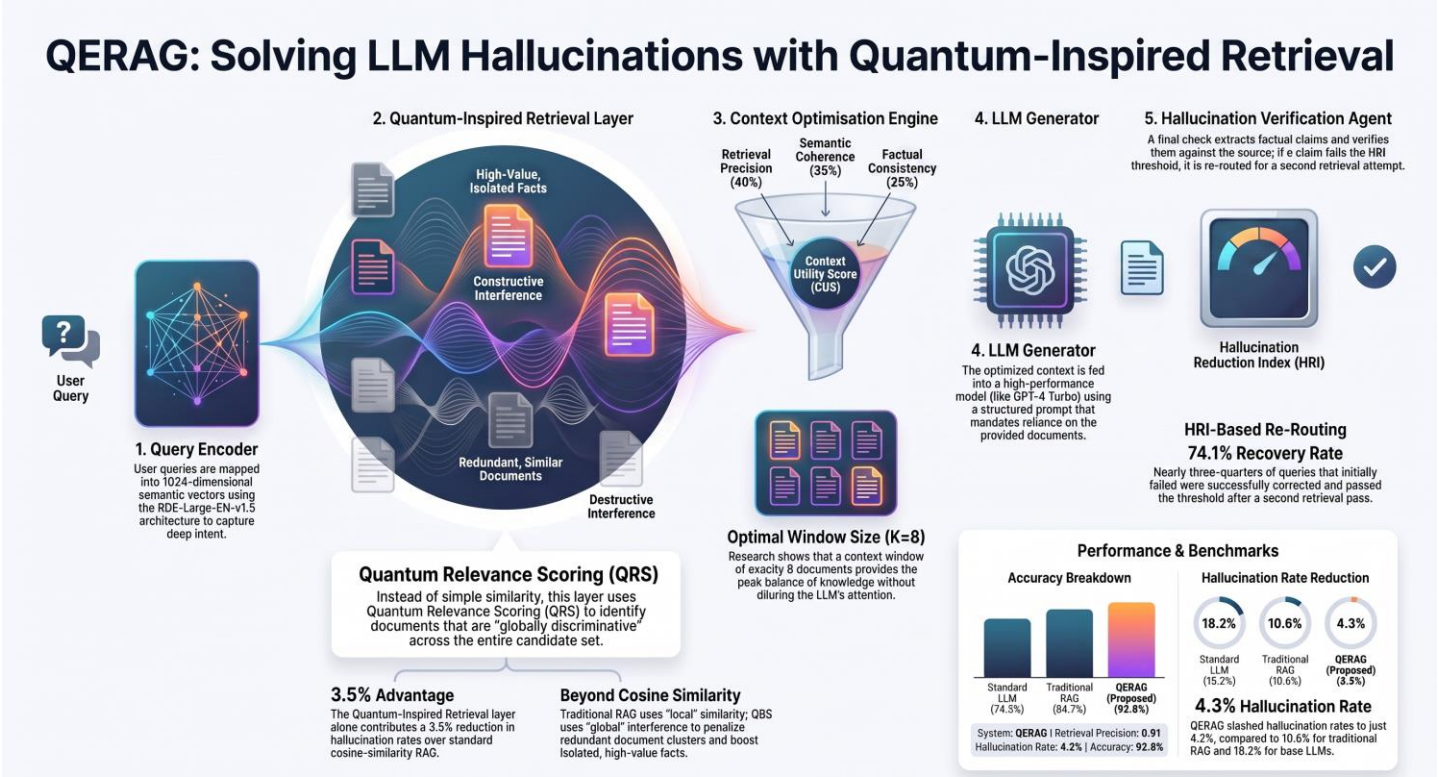
Informed Consent Statement: Not applicable.

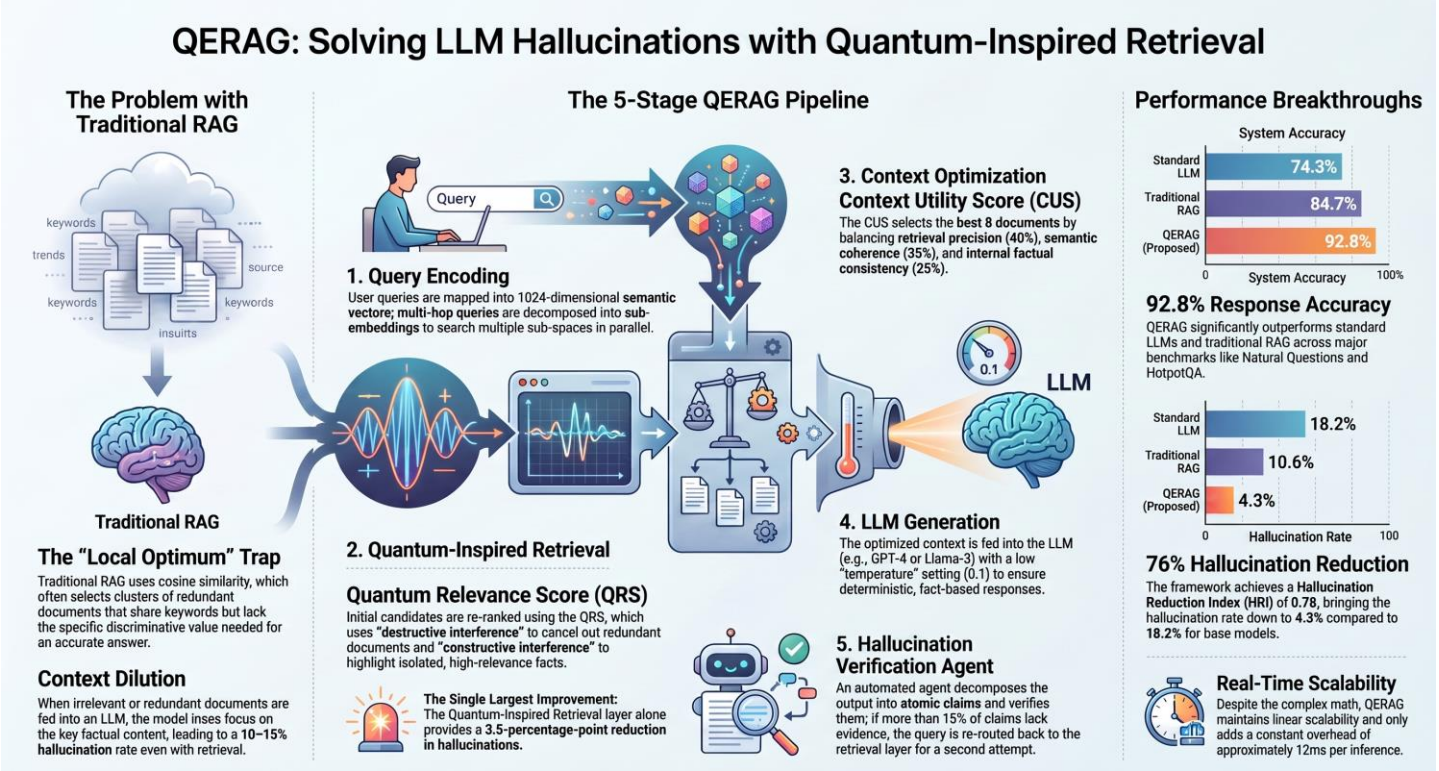
Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Acknowledgements

We thank everyone who inspired our work.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6





ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P6