electronics

MDPI

Article

Improving GPT-Driven Medical Question Answering Model Using *SPARQL*–Retrieval-Augmented Generation Techniques

Abdulelah Algosaibi [1] and Abdul Rahaman Wahab Sait [2,*]

[1] College of Computer Science and Information Technology, King Faisal University, P.O. Box 380, Al-Ahsa 31982, Saudi Arabia; aalgosaibi@kfu.edu.sa
[2] Department of Archives and Communication, Center of Documentation and Administrative Communication, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia
* Correspondence: asait@kfu.edu.sa

Abstract

The development of medical question-answering systems (QASs) encounters substantial challenges due to the complexities of medical terminologies and the lack of reliable datasets. The shortcomings of traditional artificial intelligence (AI) driven QAS lead to generating outcomes with a higher rate of hallucinations. In order to overcome these limitations, there is a demand for a reliable QAS to understand and process complex medical queries and validate the quality and relevance of its outcomes. In this study, we develop a medical QAS by integrating *SPARQL*, retrieval-augmented generation (RAG), and generative pre-trained transformer (GPT)-Neo models. Using this strategy, we generate a synthetic dataset to train and validate the proposed model, addressing the limitations of the existing QASs. The proposed QAS was generalized on the MEDQA dataset. The findings revealed that the model achieves a generalization accuracy of 87.26% with a minimal hallucination rate of 0.16. The model outperformed the existing models by leveraging deep learning techniques to handle complex medical queries. The dynamic responsive capability of the proposed model enables it to maintain the accuracy of medical information in a rapidly evolving healthcare environment. Employing advanced hallucination reduction and query refinement techniques can fine-tune the model's performance.

Keywords: medical ontologies; transformers; information retrieval; question-answering system; hallucinations; reliable healthcare systems

check for updates

Academic Editor: Ajay Bandi

Received: 7 August 2025
Revised: 28 August 2025
Accepted: 29 August 2025
Published: 31 August 2025

Citation: Algosaibi, A.; Sait, A.R.W. Improving GPT-Driven Medical Question Answering Model Using *SPARQL*–Retrieval-Augmented Generation Techniques. *Electronics* **2025**, *14*, 3488. https://doi.org/10.3390/electronics14173488

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Healthcare providers frequently encounter conditions demanding instantaneous decisions in emergency and critical care settings [1]. Traditional information retrieval techniques, including exploring textbooks, searching internet databases, or interacting with colleagues, are unsuitable for these settings [2]. Rapid access to specific medical information has a substantial impact on patient outcomes [3]. In this context, question-answering systems (QASs) serve as crucial tools, delivering prompt and precise solutions to queries and enabling faster and more informed decision-making [3]. In addition, amidst the transition towards patient-centered care models, there is an increasing focus on patient education and engagement [4]. Individuals are highly interested in learning about their health conditions and available treatment alternatives. By using QASs integrated with patient portals or mobile health applications, patients can access personalized health information. QAS enhances patient engagement and compliance with treatment strategies and reduces

the informational burden on healthcare practitioners by responding to typical individual queries [4].

The introduction of artificial intelligence (AI) in healthcare has provided an opportunity to enhance patient care and medical research [5]. However, the growth of structured and unstructured medical data poses challenges in developing an automated patient care system [6]. Unstructured data, such as medical texts, research papers, and clinical reports, offers context and information, whereas structured data, like the Resource Description Framework (RDF) standardizes data interchange [7–9]. The recent advancements in AI provide a platform for handling vast amounts of healthcare data, supporting medical professionals to treat and manage complex diseases [10].

Natural language processing (NLP) technologies, including generative pre-trained transformers (GPT), retrieval-augmented generation (RAG), and *SPARQL*, have been especially significant among the notable advancements [11]. The attention mechanisms enable GPT models to focus on the relevant part of the input sequence, capturing long-range dependencies and providing context-rich representations for each token [12,13]. Modern NLP models encounter challenges in dealing with the complexities of medical terminology, resulting in inaccurately generated content [14]. There is a lack of context-aware models that recognize and evaluate clinical communication's intricacies and unpredictability [14]. Additionally, adaptable medical NLP systems for regional- or specialty-specific word and phrase usage are currently in their early stages of development [14]. The performance of *SPARQL* relies on the RDF triples [15]. This dependency restricts it from providing valuable insights contained in unstructured data sources. A deeper understanding of the schema is crucial in formulating *SPARQL* queries. Due to these limitations, *SPARQL* queries are unsuitable in clinical decision-making environments. RAG model's performance depends on the retrieval phase.

The suboptimal information retrieval can lead to irrelevant responses. Similarly, inaccuracies and biases in the dataset can affect the performance of the GPT model [16]. This process can lead to hallucinations, generating factually incorrect information. The impact of hallucinations can be detrimental in healthcare settings, leading to serious consequences. Mitigating hallucinations can lead to the development of reliable and accurate medical QASs.

The capabilities of QASs can be considerably improved through the integration of *SPARQL*, RAG, and GPT. GPT models can provide thoughtfully designed, contextually relevant responses to multiple queries due to their proficiency in deep learning (DL) and natural language processing. Medical RDF databases typically contain clinical recommendations, research results, and pharmacological information. By leveraging *SPARQL*'s potential in retrieving precise information and RAG's capabilities of generating nuanced text from unstructured data, comprehensive, accurate, and contextually relevant responses can be generated. This synergy enables QASs to generate detailed explanations or essential context for user queries. In this study, we aim to harness the synergistic potentials of *SPARQL* and GPT, augmented by the RAG technique, to address the demands of medical data handling and query answering.

The study contributions are outlined as follows:

1. Development of QASs integrating *SPARQL* and GPT with RAG.

The novelty lies in integrating structured and unstructured data processing within a single framework. This strategy addresses a critical gap in leveraging RDG data for NLP. It enhances the reliability of generated texts by grounding them in verifiable data sources.

2. Enhancement of Data Retrieval and Processing.

By integrating sophisticated NLP techniques, this study advances the efficiency and accuracy of medical data retrieval. The innovative aspect of this contribution is the dynamic retrieval capability, extracting data based on the evolving context of an individual's health or specific clinical queries. This approach allows the proposed QASs to adapt to real-time settings, enhancing static retrieval methods.

3. Generation of a Synthetic Dataset.

In order to enable the safe and ethical training of validating AI-driven QASs, we construct a high-quality dataset mimicking real-world medical environments. This dataset serves as a benchmarking tool for evaluating QASs without risking patient confidentiality.

4. Contribution to Medical Research and Patient Care.

The proposed QAS enhances clinical decision-making and supports medical education. The integration of advanced AI techniques provides rapid, accurate medical information, aiding personalized medicine and evidence-based practice.

The study is structured as follows: Section 2 covers existing studies on *SPARQL*, LLMs, and the applications of RAG in various domains. It outlines the strengths and limitations of current information retrieval systems. Section 3 details the experimental design, data sources, and technical frameworks. It describes the creation of the synthetic dataset used for training, validation, and testing the proposed model. The quantitative outcomes, including accuracy, precision, recall, F1-Score, and the rate of hallucinations are presented in Section 4. Finally, Section 5 summarizes the key findings of the study, emphasizing the potential of the proposed QAS in handling medical queries. It reflects the broader impact of the research in improving clinical decision-making and patient care.

2. Literature Review

Advancements in QASs have been greatly facilitated by the rapid development of AI and NLP [16]. Accurate conversion of natural language queries to structured queries is essential in order to retrieve information from knowledge graphs (KGs) [17–21]. The integration of LLMs with RAG techniques has been the focus of current studies [22–24]. The primary objective of these studies is to improve the efficiency of *SPARQL* query generation and reduce hallucinations, generating non-factual information [25–27].

Ji et al. [28] enhance complex QASs through the integration of KGs into LLMs. We emphasize retrieving and reasoning capabilities, aiding in precise query generation. Muqtadir et al. [29] employ ensemble learning integrating KG and vector storage in order to mitigate hallucinations in mental health support system. Liang et al. [30] explore approaches aligning LLMs with domain-specific KGs. We enhance the natural language to graph query conversion process. Nassiri and Akhloufi [31] highlight the applications of transformers in text-based QASs. We outline the effectiveness and adaptability over multiple contexts. Badenes-Olmedo and Corcho [32] present zero-shot QASs framework using heterogeneous knowledge bases to enhance model's adaptability. Pusch and Conrad [33] mitigate the occurrences of hallucinations in QASs using the combination of LLMs with KGs. Tian et al. [34] build a dual approach to refine response accuracy and relevance. We apply RAG utilizing external knowledge.

Wu et al. [35] develop a framework based on chain-of-thought processes for better reasoning. De bellis et al. [36] explore the integration of ontologies with LLMs in order to implement effective RAG systems. Zhan et al. [37] propose an approach for retrieval-augmented reasoning on KG to enhance QAS performance in healthcare domain. Doan et al. [38] build a hybrid retrieval approach, focusing on scalability and efficiency. Xu et al. [39] introduce a unified framework leveraging LLMs with KGs to improve QAS performance. Kim (2025) [40] fine-tune LLMs with RAG to generate responses for the

medical queries. Shi et al. (2025) [41] introduce retrieval algorithms to extract medical information from structured and unstructured databases. Alonso et al. (2024) [42] train LLMs on multi-lingual datasets to handle medical queries across multiple languages. Table 1 highlights the characteristics of existing studies.

Table 1. Features and limitations of the existing studies.

Authors	Method	Features	Limitations
Ji et al. [28]	Integrating RAG and KG.	Precise query generation.	Extensive customization is required for specific KGs.
Muqtadir et al. [29]	Ensembling of KG and vector store.	Mitigating hallucination rate.	Complexities in integrating KG and vector store.
Liang et al. [30]	Domain-specific graph alignment.	Improved natural language to graph query language conversion.	Model performance is limited to specific domain.
Nassiri and Akhloufi [31]	Overview of transformer models.	Applications of transformers in QAS development.	Focused on broader applications, limiting the usage in MQAS.
Badenes-Olmedo [32]	Zero-shot QASs.	Work without specific training process.	May face challenges in processing medical queries.
Pusch and Conrad [33]	Integration of LLMs with KG.	Reducing hallucinations.	Emphasizing QAS accuracy rather than context.
Tian et al. [34]	RAG with external knowledge.	Applying external knowledge for model performance improvement.	Performance of the model relies on quality of external knowledge.
Wu et al. [35]	Chain-of-thought in KG for QAS development.	Integration of reasoning steps.	Demand for sophisticated modeling.
De Bellis et al. [36]	Integrating ontologies and LLMs.	Enhancing RAG by applying ontological knowledge.	Challenges in managing ontologies.
Zhan et al. [37]	Effective framework for medical QASs.	Specialized for healthcare domain.	Limited generalization ability.
Doan et al. [38]	Hybrid retrieval capability.	Enhancement of scalability and reliability.	Complexities in model implementation.
Xu et al. [39]	KG-based QASs.	Adaptive multi-aspect RAG.	Limited discussion on hallucination mitigation.
Kim [40]	Fine-tuned LLMs with RAG.	Tailored optimizations to handle medical terminologies.	Model's performance relies heavily on the quality of data sources.
Shi et al. [41]	Specialized retrieval algorithm for medical information extraction.	Incorporating feedback mechanisms for improving retrieval and generation over time.	Integrating retrieval and generation components may lead to increased system complexity.
Alonso et al. [42]	LLMs with RAG.	Multi-lingual QAS.	Medical terminologies vary significantly across regions, affecting the applicability of responses across different settings.

The findings of these studies underscore the importance of integrating LLMs with RAG approaches and structured data retrieval methods, including *SPARQL*, in order to improve the accuracy and reliability of QASs in specialized domains. The current QASs using *SPARQL*, RAG, and LLMs exhibit significant progress. However, future enhancement is required to address gaps in integrating structured and unstructured data sources. Although *SPARQL* is capable of effectively querying structured databases, its integration with deep learning models to improve the accuracy and relevance of results obtained from unstructured data is underexplored. In healthcare settings, existing models face challenges in retrieving relevant data and providing contextually accurate responses. As a result, challenges like information retrieval bottlenecks and retrieving irrelevant data arise, leading to user dissatisfaction. There has been a lack of effective strategies to identify and reduce hallucinations in generative models. There is a demand for extensive research on the interaction between hallucination tendencies, training data quality, and model architecture.

3. Materials and Methods

In this study, we enhance the functionality and efficiency of medical information systems through the integration of NLP and advanced AI techniques. The study focuses on the transformative potential of RAG and the sophisticated querying potential of *SPARQL* within RDF-based knowledge graphs. The transformation of natural language queries to *SPARQL* serves as the gateway to interact with structured knowledge bases. The application of the state-of-the-art transformer architectures can improve the accuracy and reliability of medical information systems. This approach minimizes hallucination effects in order to generate factually correct information. Figure 1 shows the synergy of RDF triples, RAG-based document retrieval, and synthetic data generation.

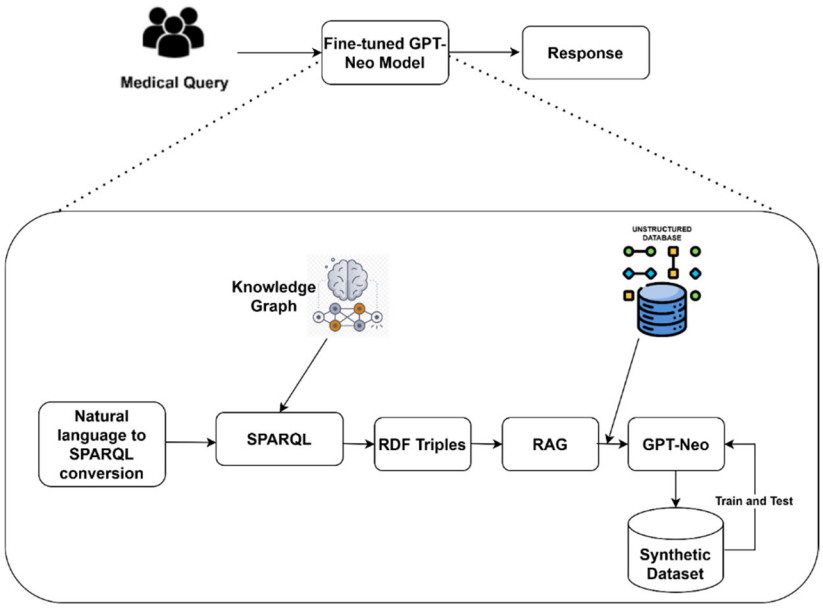

Figure 1. The proposed approach for answering the user queries.

3.1. Data Acquisition

In this study, we employ medical ontologies, including SNOMED CT [43], LOINC [44], MeSH [45], RdfNorm [46], UMLS [47], and Gene Ontology [48], to retrieve RDF triples, providing the foundation data needed for accurate responses. These ontologies ensure that the information is comprehensive and standardized across diverse platforms and systems. To retrieve unstructured data, we utilize key repositories, including PubMed [49], Clinical

trials [50], RadioPedia [51], and Open-I [52]. In addition, to evaluate the model's generalization capability, we use the MEDQA dataset [53] covering 766 question–answer pairs.

3.2. Natural Language to SPARQL Transformation

The conversion of natural language queries to *SPARQL* plays a crucial role in performing semantic searches on RDF data stores. It retrieves semantically relevant information, addressing the challenges in understanding the relationship between symptoms, diseases, and interventions. To ensure the compatibility between the query and the structure and semantics of data in the knowledge graph, we apply query optimization and validation strategies. *SPARQL* is a standardized query language, facilitating simplified data sharing and interpretability. It can extract specific data relevant to patient characteristics, contributing to personalized treatment plans. Equation (1) highlights the suggested transformation strategy.

$$Q_{SPARQL} = f_{NLP}(Q_{NL}, K_{Schema}) \tag{1}$$

where Q_{SPARQL} is the *SPARQL* query, and f_{NLP} is the NLP function.

The f_{NLP} function interprets and maps user queries (Q_{NL}) based on the defined schema (K_{Schema}). The *SPARQL* formulation leverages the underlying ontology of the knowledge graph, providing solid foundation for the subsequent processes in the proposed medical information retrieval pipeline.

3.3. RDF Triple Retrieval

In the proposed methodology, the RDF triple retrieval serves as the bridge between structured knowledge representation and practical information application. It guarantees the accuracy and reliability of the extracted responses. Equation (2) reveals the process of executing the query to retrieve appropriate RDF triples from the knowledge graph.

$$R_{RDF} = KG(Q_{SPARQL}) \tag{2}$$

where R_{RDF} is the RDF triples extracted from the knowledge graph (KG).

The KG function uses Q_{SPARQL} to extract the RDF triples using the knowledge graph. The extracted RDF triple comprises subject–predicate–object constructs, encapsulating structured information. For instance, the query "diabetic symptoms" may return triple like (Diabetes, hasSymptom, FrequentUrination). The RDF triple enriches the context generation process. The query optimization and validations support the proposed approach in mapping *SPARQL* queries to the relevant triples, reducing the risk of irrelevant or inaccurate data.

3.4. RAG-Based Document Retrieval

To enrich the structured data points (RDF triples), we employ the RAG framework. By capitalizing on the semantic richness embedded within the RDF triples, RAG conducts targeted searchers across document repositories, including research papers, clinical guidelines, and comprehensive medical articles. Equation (3) presents the mathematical expression of the RAG-based document retrieval.

$$D_R = f_{RAG}(R_{RDF}, C_{doc}) \tag{3}$$

where D_R is the retrieved documents using the function (f_{RAG}) from the corpus (C_{doc}).

The retrieval process is fine-tuned to the semantic nuances of the RDF data, guiding the retrieval mechanism towards documents containing relevant information. The f_{RAG} function generates a semantic query set. We compose canonical labels and synonyms for the subject/object entities, relation-aware cues derived from the predicate, and salient

qualifiers present in the graph, populating the semantic query set. The semantic query set is transformed into dense vectors with the Sentence_BERT. To minimize the computational costs, principal component analysis is used. The cosine similarity technique is used to evaluate the similarity between the query and the extracted documents' vectors. The extracted documents are re-ranked by combining similarity with structured features, including entity coverage, predicate alignment, qualifier match, and source trust. We de-duplicate near-duplicates, cap per-publisher dominance, and retain provenance for auditability. The final set of documents are passed to the generator for evidence-gated answering. During response generation, we employ the permutation importance technique to compute the feature importance value for each feature during the response generation for a query.

3.5. GPT-Neo Outcome Generation

GPT-Neo [54] provides a robust, scalable, and adaptable solution, capable of handling complex queries and mitigating the risk of hallucinations. It is an open-sourced and optimized model for democratized access. Researchers and developers can tweak the GPT-Neo's architecture, training procedure, and data processing pipelines. GPT-Neo's scalability enables developers to maintain trade-offs between performance and computational costs. By leveraging a highly curated dataset, the occurrence of hallucinations can be decreased. These features motivate us to employ GPT-Neo to generate an outcome using RDF triples and unstructured data. GPT-Neo integrates and synthesizes RDF triples and unstructured data into an informative outcome. Equation (4) presents the process of output generation.

$$O_N = f_N(D_R, \theta) \tag{4}$$

where O_N is the GPT-Neo's output using the function (f_N) and θ is the GPT-Neo's parameters.

The f_N function transforms the retrieved documents (D_R) into an outcome (O_N). The generated outcomes (O_N) are based on the combined information from RDF triples and associated unstructured documents.

3.6. Synthetic Dataset Generation

The pre-trained GPT-Neo may face challenges in handling real-time medical data. It may generate misleading information due to the complexities of medical queries. To overcome the shortcomings, high-quality annotated synthetic data is required. Generating synthetic data exposes GPT-Neo to a diverse range of scenarios. The trained GPT-Neo can support evidence-based medicine due to its capability of synthesizing datasets containing the latest research and clinical findings. We generated queries using medical ontologies (SNOMED CT, MeSH, UMLS, and Gene Ontology) through sampling of entities/relations and composing *SPARQL* to retrieve RDF triples. These triples are seeded RAG, retrieving top-K passages from PubMed, ClinicalTrails.gov, RadioPedia, and Open-i. GPT-3 generates question–answer pairs conditioned on the triples and retrieved evidence. We apply pre-processing techniques, including tokenization, normalization, and data encoding into an appropriate format. GPT-Neo uses the pre-processed dataset to generate detailed and contextually informed answers. These outcomes are paired with the original query to form the synthetic dataset, comprising paired question–answer entries. Equation (5) outlines the computational form of synthetic dataset generation.

$$D_S = \left\{ \left(Q_{NL}^i, O_N^i \right) \right\}_{i=1}^{N} \tag{5}$$

where D_S is the synthetic dataset and N is the total number of queries and outcomes, Q_{NL}^i is the user query, and O_N^i is the generated outcome.

We used scheduled scripts to automatically update the dataset without manual intervention, ensuring seamless integration of novel data. Using this approach, the dataset remains up-to-date with the latest medical advancements and discoveries. In addition, we enable the model to gather feedback from users through user correction inputs and analysis of follow-up questions, indicating confusion or dissatisfaction with previous responses. The synthetic dataset is balanced, audited, and validated in order to enable reliable use of the dataset in the healthcare domain. The process of stratifying along three dimensions, including disease families (mapped to ICD/SNOMED chapters), question intent (diagnostic, treatment, medication, risk, prognosis, prevention, monitoring, complications, lifestyle, education), and difficulty tier, is used to ensure that balancing is properly implemented at the generation time. Each item's difficulty depends on *SPARQL* graph depth, number of joins, and RAG's document volume and heterogeneity. Using a stratified approach, two domain experts examined and annotated the synthetic dataset for factual correctness, clinical appropriateness, clarity, and safety. Reviewers are unaware of the generator or the prompts in order to reduce the possibility of bias. In the event that no agreement can be achieved following a series of brief adjudication sessions, a third senior clinician makes the final decision. We measure consistency through inter-rater agreement and breakdown revision types: minor language modifications, evidence updates, clinical corrections, and discards. Additionally, we normalize entities to SNOMED/MeSH and drugs to generics in order to prevent brand or region effects. To generalize the model on the external dataset, we utilize the MEDQA dataset [44].

3.7. Fine-Tuning GPT-Neo

We train GPT-Neo using the synthetic dataset with Bayesian Optimization and Hyperband (BOHB) optimization. The model learns the nuances of medical terms and the structure of medical knowledge. BOHB determines the optimal learning rates, batch sizes, and other key parameters influencing the model's performance. It supports GPT-Neo to understand and integrate the context, minimizing the risk of hallucinations. During fine-tuning, the loss function is used to measure the discrepancy between the generated responses and the actual responses in the dataset. Let x_i be the tokenized input for the ith instance in the synthetic dataset. This input includes query and relevant context. Equation (6) shows the functionality of GPT-Neo, representing the operational mechanism of GPT-Neo during the fine-tuning process.

$$f_N(x_i, \theta) \tag{6}$$

where θ is the GPT-Neo's hyperparameters.

The f_N function produces a sequence of logits over the possible output tokens in the model's vocabulary. Equation (7) computes the cross-entropy loss between the predicted probability distribution and the actual distribution of the subsequent tokens in the training set.

$$L(\theta) = -\sum_{j=1}^{n} \log P\left(y_{i_j} \middle| x_i, y_i < j; \theta\right) \tag{7}$$

where L is the loss function, P is the probability distribution, y_i is the output token (response), $y_i < j$ is the sequence of the actual tokens before position j, and i is the input instance.

We employ top-k sampling approach to enhance the quality of the generated text during the inference phase. This approach uses top-k words with the highest probability. It

introduces randomness into the response generation, producing diverse and meaningful response. Equation (8) presents the generated response.

$$y_{i_j} \sim P_k(\omega | x_i, y_i < j; \theta) \tag{8}$$

where y_{i_j} is the generated outcome during the inference phase, P_k is the probability distribution over the vocabulary, and ω is a token in the vocabulary.

3.8. Performance Evaluation

The evaluation metrics, including accuracy, precision, recall, and F1-score are employed to evaluate the effectiveness of the proposed model. These metrics assess the quality of the generated responses, ensuring the model's reliability. There is no standard hallucination metric for evaluating the generated text [55–57]. Equation (9) is computed as the mean of a binary indicator over the evaluation set. Averaging these indicators yields a proportion between 0 and 1, quantifying the existence of hallucinated answers across the corpus. We breakdown the text into checkable claims and assemble an evidence pool for each model response. The evidence pool comprises *SPARQL*-verified RDF triples and the top-K passages retrieved by the RAG component. The model evaluates each claim against this pool, yielding an outcome. Using a clinician-adjudicated set, decision thresholds for support and contradiction are calibrated. Evaluation reports the answer-level hallucination rate over the number of retrieved documents, aligning this metric with the safety requirements for reliable MQAS.

$$\begin{aligned} H = \frac{1}{N} \sum_{i=1}^{N} 1 \Big[\ni j : \; & maxPr\Big(entail \Big| e \to C_{i_j}\Big) \\ & < \tau_{sup} \bigvee maxPr\Big(contradict \Big| e \to C_{i_j}\Big) \\ & > \tau_{con} \bigvee maxPr\Big(contradict \Big| t \to C_{i_j}\Big) > \tau_{KG} \Big] \end{aligned} \tag{9}$$

where H is the rate of hallucination rate; N is the total number of evaluated answers; $\ni j$ is a claim; C_{i_j} is the j-th atomic claim extracted from answer i; max is the best maximum score over all evidence items; $Pr(.)$ is the function that returns probabilities between 0 and 1; τ_{sup} , τ_{con}, and τ_{KG} are the thresholds for support, contradictions, and knowledge graph (KG); and $\bigvee$ is the logical OR function.

We employ statistical methods, including 95% confidence intervals (CI) and p-values to support the model's outcomes. The 95% CI expresses the uncertainty around the estimated metrics (accuracy and hallucination rate). Practically, 95% corresponds to standard errors in normal-based intervals with the common $p < 0.05$ significance level.

4. Results and Discussion

We utilize Windows 11 Pro with Intel i9 (Intel, Santa Clara, CA, USA) and an NVIDIA H100 GPU (NVidia, Santa Clara, CA, USA) to ensure compatibility and stability throughout the model implementation process. The synthetic dataset contains 6500 question–answer pairs while the MEDQA dataset covers 766 question-answer pairs. The dataset is divided into training (60%), validation (20%), and testing (20%). In addition, a significant portion (20%) of the MEDQA dataset is used to evaluate the model's generalization, guaranteeing the model's capability on novel data. The internal and external evaluation comprises 1453 inputs and outputs (1300 internal test and 153 external test). Python 3.10.0 is used for the robust and efficient implementation of the model, integrating *SPARQL*, GPT-Neo, and the RAG framework. The key Python 3.12 libraries include Transformers 4.10.0, RDFlib 5.0.0, Pandas 1.3.3, AllenNLP 2.9.3, MatPlotlib 3.4.3, Seaborn 0.11.2, TensorFlow 2.6.0, and

Optuna 2.10.0. Table 2 offers the details of the computational strategies for deploying the proposed model.

Table 2. Model implementation environment.

Component	Specification
GPT-Neo	2.7 B-Shinen
Epochs	80
Learning Rate	1×10^{-5}
BOHB Optimization	10 Iterations
SPARQL Endpoint	Local RDF store
Document Retrieval	Fine-tuned for medical context
RAG	Integrated with transformers
Dynamic Adjustment	Feedback loops for continuous improvement

Figure 2 illustrates an end-to-end instance of transforming a medical query into a detailed and useful response, aiding patients and healthcare providers in comprehending diabetes treatment options. Each step builds upon the previous one in order to ensure that the outcome is reliable and grounded in verified medical knowledge.

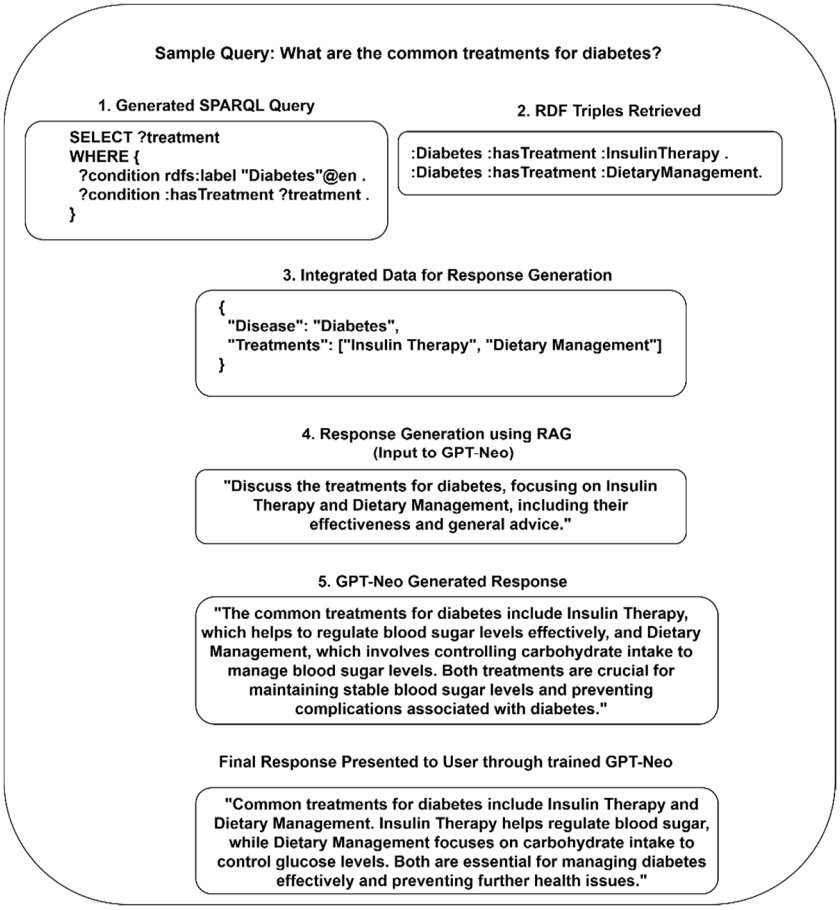

Figure 2. Process of transforming a medical query into a detailed response.

Figure 3 shows the training and validation accuracy of the proposed model across epochs. Using parameter-efficient tuning, we fine-tuned the GPT-Neo model. By freezing

the base model weights, the GPT-Neo model was trained. We focus LoRA adapters on the transformer blocks, LayerNorm gains, biases, and the LM-head, during the training phase. BOHB optimization searched the hyperparameters, including lr, weight decay, warmup ratio, LoRA rank/α/dropout, batch size, maximum sequence length, and scheduler. In order to validate the model's resilience in real-world situations, an early stopping strategy is used to ensure that the model maintains its generalizability to unseen data. The precision of the training exhibits steady improvement beginning with the first epoch, extending significantly until it reaches a plateau around the 25th epoch. This increasing trend shows that the model learns from the training data efficiently and adjusts its parameters to generate the outcomes. A robust fit to the training data implies that there are no apparent indicators of underfitting, as the training accuracy reaches its maximum level (85%). Implementing the early stopping strategy at the 47th epoch is beneficial. This approach halts training when validation accuracy no longer improves, avoiding overfitting. While retaining model generality, the early stopping point is selected to maximize validation accuracy.

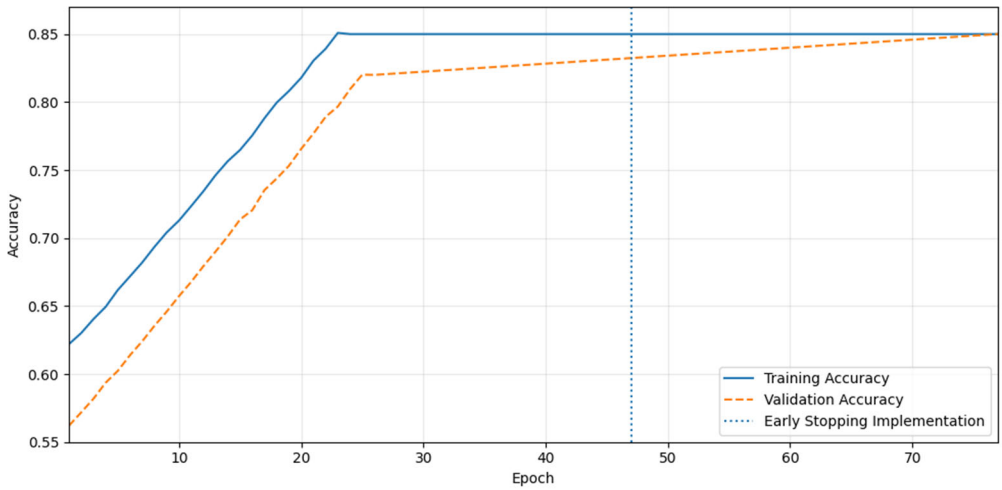

Figure 3. Performance of the proposed model during training and validation phase.

A comprehensive performance analysis of the proposed model on the synthetic dataset is presented in Table 3. The analysis covers five distinct sets and includes parameters such as accuracy, precision, recall, F1-score, and hallucination rate. The consistency in accuracy and precision throughout the majority of the sets indicate the exceptional performance. With significant documents retrieved, recall increases to 90.2% in Set 5, retrieving appropriate documents. The model retrieves relevant documents with high accuracy and low false positives, which is crucial for QAS dependability. With a balance between accuracy and recall, the F1-Score displays an amazing peak at 90.3% in Set 5, underlining the model's efficiency in balancing recall and accuracy at larger document retrieval volumes. In Set 5, the hallucination rate of 0.25 indicates that the system is becoming effective as more documents are retrieved. Due to early stopping strategies, the average test accuracy (85.7%) of the proposed model is slightly higher than the training accuracy (85%) This trend can indicate the potential of the proposed QAS in generating more accurate and less hallucinated responses.

Table 4 demonstrates the individual and combined contributions of various components in the proposed model. This analysis is used to benchmark the efficacy of each module within the model's architecture. The baseline model achieves an accuracy of 85.7% with F1-score of 87.5%. The relatively low hallucination rate of 0.25 underscores the model's reliability in generating accurate and relevant responses. Excluding the *SPARQL* retrieval led to the increased hallucination rate, suggesting the significance of *SPARQL*

retrieval in producing factual content. The removal of the RAG component reduces the model's accuracy of 78.4% with a hallucination rate of 0.45. The inherent features, including parametric knowledge and template/lexical regularities, supported the model to achieve a considerable outcome without *SPARQL* and RAG. The outcome indicates the RAG's role in enhancing the quality and depth of the generated answers. Without hallucination checks, the model obtains 89.8%, suggesting hallucination checks prevent the generation of incorrect information. Finally, the model's performance is significantly reduced without *SPARQL*, RAG, and hallucination checks. Thus, the absence of *SPARQL* and LLMs with RAG, lead to a substantial increase in error rates and ineffective handling of queries.

Table 3. Findings of performance analysis—synthetic dataset.

Set	Number of Documents Retrieved	Relevant Documents	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Hallucination Rate
1	10	9	90.0	90.1	80.2	84.8	0.24
2	20	17	85.0	90.2	85.4	87.7	0.27
3	30	26	86.6	84.1	87.3	85.6	0.23
4	40	34	85.0	87.8	90.6	89.1	0.28
5	50	41	82.0	90.3	90.3	90.3	0.25
Average	30	25.6	85.7	88.5	86.7	87.5	0.25

Table 4. Findings of ablation study.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Hallucination Rate
Proposed Model (Baseline)	85.7	88.5	86.7	87.5	0.25
Without *SPARQL* and RAG	78.4	80.6	80.4	80.5	0.45
Without Hallucination checks	89.8	80.3	77.9	79.0	--
Without *SPARQL*, RAG, and Hallucination checks	62.1	59.26	61.02	60.1	--

Table 5 presents a comparative performance analysis of various transformer-based QASs against the proposed QAS on a synthetic dataset. The baseline models are underperformed due to the lack of KG-aware re-ranking and evidence-gated decoding. T5 demands more context or longer outputs to resolve contraindication details. The compressed representation affects the performance of DistilBERT, generating the highest hallucination rate (0.67 ± 0.018) and modest F1-Score. With fixed prompt templates, XLNet produces generic response. However, the lack of KG-aware generation restricts its performance. The low hallucination rate of 0.25 underscores the model's ability to generate responses with minimal errors. The superior performance can be credited to the recommended *SPARQL* and GPT-Neo integration. The findings highlight the effectiveness of the proposed model in producing meaningful responses to the user queries.

Figure 4 illustrates the model's performance on the MEDQA dataset based on different volumes of documents. It indicates that the increased number of documents influence the model's performance. The high precision value suggests the effectiveness of the model in identifying relevant responses. The high F1-score highlights an overall improvement in balanced performance, rendering robust scalability in handling an increasing number of documents.

Table 5. Findings of comparative analysis (different transformers on synthetic dataset).

Model	Accuracy (%) (95% CI)	Precision (%)	Recall (%)	F1-Score (%)	Hallucination Rate (95% CI)	GPU (Seconds/Query)
Proposed Model	85.7 ± 1.6	88.5	86.7	87.5	0.25 ± 0.017	0.35
BERT	78.1 ± 1.9	74.4	72.8	73.5	0.39 ± 0.019	0.61
RoBERTa	73.6 ± 2.1	70.5	71.3	70.9	0.48 ± 0.020	0.63
T5	79.6 ± 1.7	75.3	73.8	74.5	0.56 ± 0.019	0.87
DistilBERT	81.2 ± 1.5	79.5	77.1	78.2	0.67 ± 0.018	0.58
XLNet	81.6 ± 1.6	78.7	75.7	77.1	0.47 ± 0.019	0.49

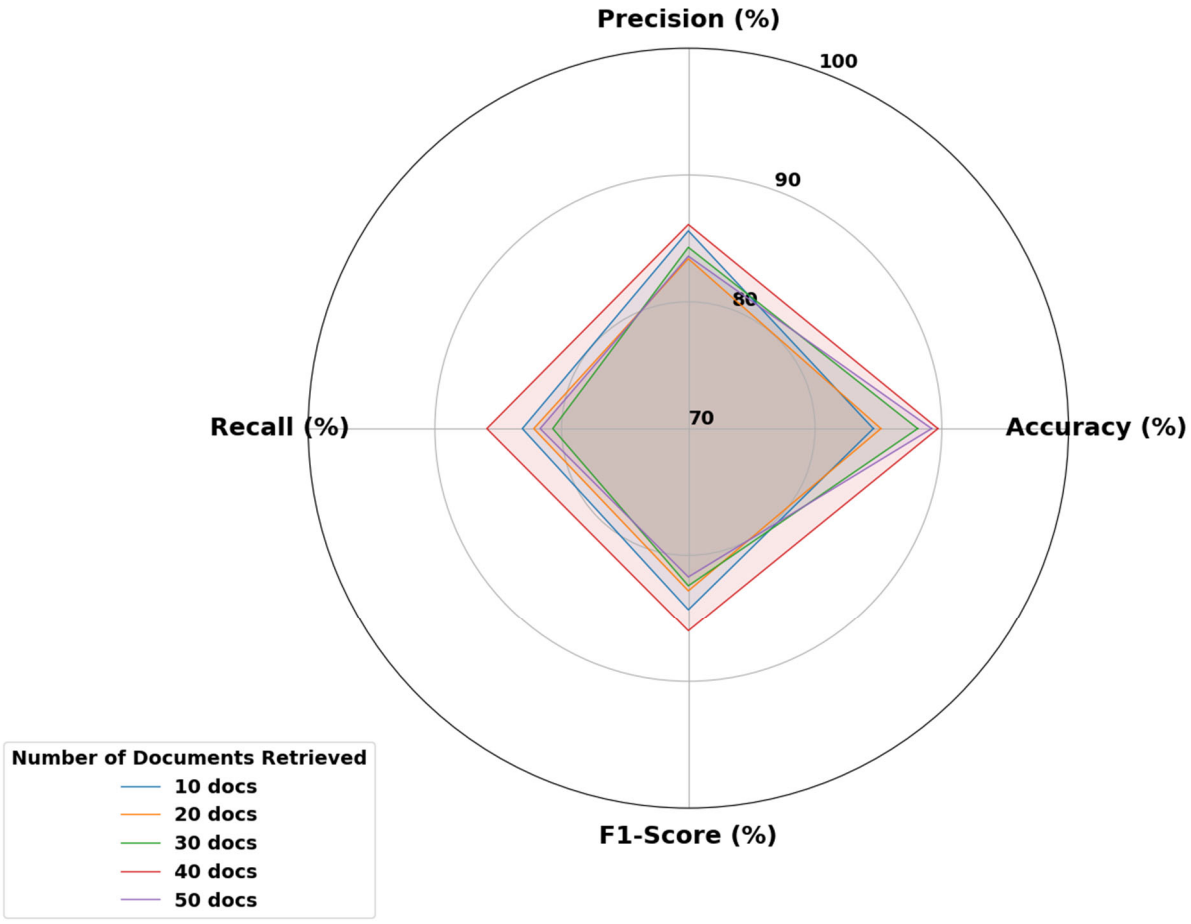

Figure 4. Findings of performance analysis of the proposed model on MEDQA dataset.

Table 6 outlines a comparative analysis of the proposed QAS against well-known transformers using the MEDQA dataset. The proposed model shows substantial improvements compared to its counterparts. It identifies relevant information and minimizes incorrect information, which is a key factor in medical contexts. We used bucketing/padding to a constant sequence length. In addition, the caches are warmed for *SPARQL* and the vector store. Under this environment, the GPU time is nearly identical across datasets. The proposed model reports a hallucination rate of 0.31 ± 0.020, significantly lower than other models, highlighting its superiority in generating reliable responses. We report 95% CI and evaluate pairwise gaps at a significance level of $p < 0.05$. Paired significance tests on per question outcomes confirm that the proposed model surpasses the baseline models with $p < 0.05$. We adjust for multiple comparisons, and the outcomes remain significant at

$p < 0.05$. The superior performance implies the model's potential integration into clinical practice, supporting clinical decision-making, patient education, and engagement.

Table 6. Findings of comparative analysis (different transformers on MEDQA dataset).

Model	Accuracy (%) (95% CI)	Precision (%)	Recall (%)	F1-Score (%)	Hallucination Rate (95% CI)	GPU (Seconds/Query)
Proposed Model	87.36 ± 1.8	84.60	82.72	83.65	0.31 ± 0.020	0.35
BERT	71.27 ± 2.3	64.46	63.67	64.06	0.49 ± 0.025	0.61
RoBERTa	69.48 ± 2.5	72.55	69.85	71.17	0.54 ± 0.019	0.63
T5	70.72 ± 3.1	71.39	70.43	70.91	0.46 ± 0.023	0.87
DistilBERT	71.33 ± 2.4	68.96	70.31	69.63	0.53 ± 0.031	0.58
XLNet	71.79 ± 3.4	70.11	71.26	70.68	0.59 ± 0.024	0.49

Figure 5 illustrates the area under the receiver operating characteristics (AUROC) curves for the model on the synthetic and MEDQA datasets. An AUROC of 0.93 on the synthetic data represents the model's discriminatory ability. The high AUROC of 0.91 on the MEDQA dataset indicates the model's robust performance on real-world data. The outcomes present compelling evidence to show the capabilities of the proposed QAS in generating accurate and reliable responses to medical queries.

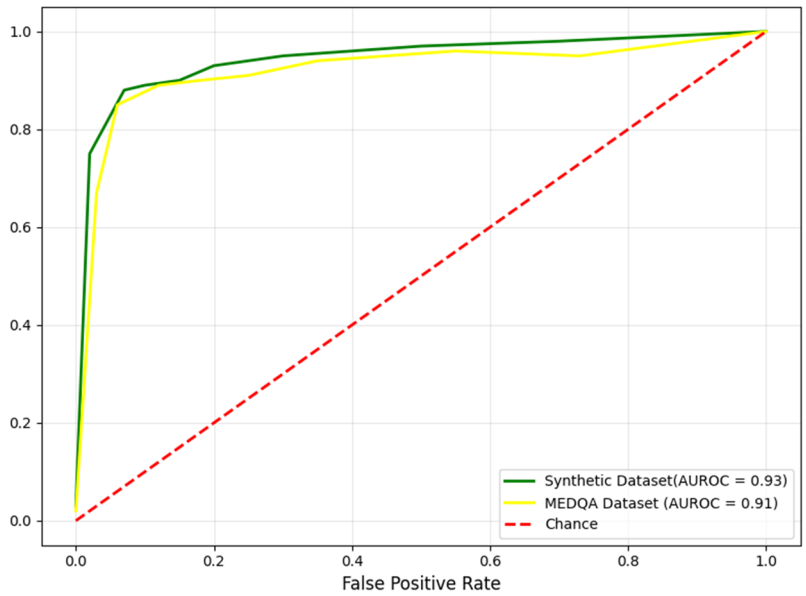

Figure 5. Findings of AUROC analysis using synthetic and MEDQA datasets.

Table 7 reveals a comparative analysis of various models, including the proposed QAS, across different datasets. Shi et al. [41] and Alonso et al. [42] generate responses with high hallucination rates, reflecting limited grounding or weaker evidence controls. Kim (2025) [40] exhibits higher accuracy with substantially higher hallucination rate. In contrast, the proposed model generates a response with minimum hallucination rate of 0.31 ± 0.020 on the MEDQA dataset, demonstrating strong generalizability and effectiveness in real-world settings. The incorporation of adaptive learning techniques refine the proposed model's performance on novel data, maintaining its reliability with the latest medical information and user interaction trends.

Table 7. Findings of comparative analysis (proposed model against existing state-of-the-art models using the MEDQA dataset).

Model	Accuracy (%) (95% CI)	Precision (%)	Recall (%)	F1-Score (%)	Hallucination Rate (95% CI)	GPU (Seconds/Query)
Proposed Model	87.36 ± 1.8	84.60	82.72	83.65	0.31 ± 0.020	0.35
Shi et al. [41]	61.45 ± 2.5	59.65	61.24	60.43	1.05 ± 0.018	0.87
Zhan et al. [37]	71.23 ± 1.2	69.84	65.34	67.52	0.63 ± 0.025	0.75
Alonso et al. [42]	75.84 ± 1.6	71.45	69.64	70.53	1.12 ± 0.019	0.64
Xu et al. [39]	80.21 ± 2.1	79.63	76.67	78.12	0.89 ± 0.026	0.55
Kim [40]	86.66 ± 1.2	77.36	75.04	76.18	0.95 ± 0.021	0.48

The experimental findings indicate that the proposed model outperforms current QASs and transformers-based information retrieval systems. Compared to individual transformers, the proposed QAS generate responses with minimal rate of hallucinations. By leveraging the hybrid *SPARQL*–RAG–GPT-Neo approach, the model handles complex language patterns and long-range dependencies within texts. The integration of advanced technologies improve information retrieval, decision-making, patient interaction, and clinical operations, advancing healthcare technology. Training on a broad spectrum of medical text and datasets, we enabled the model to comprehend standard medical queries and lesser-known medical terminologies. This extensive training approach facilitates the model to manage a wide variety of medical queries, reducing biases and improving generalization.

In order to improve the accuracy of patient diagnosis and treatment, the proposed model can assist healthcare professionals by properly understanding complicated medical queries and obtaining relevant information from extensive healthcare databases. For instance, a clinician may retrieve recent treatment procedures for chronic conditions, including diabetes or hypertension. To deliver evidence-based suggestions, the model can process the latest research, clinical guidelines, and patient data. This reduces duration and guarantees that treatment is based on the latest medical knowledge, increasing patient outcomes. As virtual health assistants, the model can respond to health-related queries, clarify prescription regimens, and provide guidance on minor ailments. It empowers users to control their health, improving health outcomes and increasing their satisfaction. The model can identify a patient's requirements and provide recommendations for treatment and prevention based on their medical history, genetic profile, and lifestyle habits. By recommending lifestyle adjustments and preventative care targeted to individual health requirements, it improves treatment effectiveness and disease prevention. Improved diagnosis accuracy and treatment effectiveness enable the model to contribute to lowering the overall healthcare cost. Stopping misdiagnoses and unsuitable treatments saves a lot of money and lessens the financial load on patients and healthcare providers. This approach can assist researchers in accelerating the pace of medical research and discoveries by facilitating the development of hypotheses, the design of experiments, and the analysis of study data.

Despite its strengths, the study acknowledges certain limitations, restricting its generalizability and scalability. Using a synthetic dataset for controlled experiments may be unable to adequately replicate the complexity and diversity of real-time data, especially in medical settings where queries can be unpredictable. Although the model performs extremely well, its accuracy decreases as query complexity expands, indicating the demand for further enhancements in the model's architecture. Using cutting-edge machine learning methods, including meta-learning and few-shot learning, may enable the model to adapt rapidly to complex queries without extensive retraining. The clinical relevance and usefulness of the

model can be improved by iterative refinement in collaboration with medical experts based on practical input. This would guarantee that the model closely matches clinical goals and priorities in addition to performing effectively. Investigating additional factors, including more sophisticated methods of reducing hallucinations and recent NLP advancements can improve the model outcomes, guaranteeing dependability and reliability.

Prioritizing regulatory, ethical, and patient-safety measures is essential for validating the proposed model using real-time clinical data. Essential controls include evidence-gated generation (answers traceable to referenced knowledge-graph triples and retrieved sources); rigorous safety checks for dosages, contraindications, and black-box warnings; and conservative abstention with explicit uncertainty messaging when evidence is inadequate. Privacy measures should adhere to existing health data regulations, which include data minimization and the implementation of robust access controls. Lifecycle governance should include post-release bias and safety monitoring, incident reporting, and restricted knowledge graph and model updates. Prior to any clinical review, human-factors testing, risk management associated with medical device practices, and evaluation of AI applications, are essential. These strategies will provide a visible, auditable path from research prototype to clinically ideal applications.

5. Conclusions

This study establishes a robust foundation for medical informatics, offering an effective tool to enhance clinical decision-making and patient outcomes. The innovative integration of advanced NLP techniques, including *SPARQL*, KG, GPT-Neo, and RAG, enabled the proposed model's capabilities to evolve with the rapidly changing landscape of medical research. The primary contribution of this study is its ability to reduce the hallucinated responses, an existing challenge in AI-driven QASs. The dynamic design of the model emphasizes scalability and reliability, handling an expanding scope of medical queries without compromising model's performance. Moreover, the dynamic adaptability guarantees that the model remains relevant and up-to-date, reflecting the latest medical standards and practices. However, the complexity of integrating *SPARQL*, GPT-Neo, and RAG demands substantial computational resources. Although the model has a lower rate of hallucinations, it may produce outcomes with occasional inaccuracies to queries with less common medical conditions or treatments. There are several future avenues for further enhancement. The model's capability can be refined with broader training datasets covering rare diseases and treatments. Another potential development is to enhance the model's language capabilities to handle multiple languages, allowing the model to extend its services across the globe.

Author Contributions: Conceptualization, A.A. and A.R.W.S.; Methodology, A.A. and A.R.W.S.; Software, A.A. and A.R.W.S.; Validation, A.A. and A.R.W.S.; Formal analysis, A.A. and A.R.W.S.; Investigation, A.A.; Resources, A.A. and A.R.W.S.; Data curation, A.R.W.S.; Writing—original draft, A.R.W.S.; Writing—review & editing, A.A.; Visualization, A.A. and A.R.W.S.; Supervision, A.R.W.S.; Project administration, A.R.W.S.; Funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU253047].

Data Availability Statement: The dataset is available at https://huggingface.co/datasets/bigbio/mediqa_qa (accessed on 31 January 2024).

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

References

1. Zhu, Y.; Wang, X.; Chen, J.; Qiao, S.; Ou, Y.; Yao, Y.; Deng, S.; Chen, H.; Zhang, N. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web* **2024**, *27*, 58. [CrossRef]
2. Zhang, Y.; Sui, X.; Pan, F.; Yu, K.; Li, K.; Tian, S.; Erdengasileng, A.; Han, Q.; Wang, W.; Wang, J.; et al. A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. *Nat. Mach. Intell.* **2025**, *7*, 602–614. [CrossRef]
3. Jolfaei, S.A.; Mohebi, A. A review on persian question answering systems: From traditional to modern approaches. *Artif. Intell. Rev.* **2025**, *58*, 127. [CrossRef]
4. Perevalov, A.; Both, A.; Ngonga Ngomo, A.C. Multilingual question answering systems for knowledge graphs—A survey. *Semant. Web* **2024**, *15*, 2089–2124. [CrossRef]
5. Qu, C.; Dai, S.; Wei, X.; Cai, H.; Wang, S.; Yin, D.; Xu, J.; Wen, J.R. Tool learning with large language models: A survey. *Front. Comput. Sci.* **2025**, *19*, 198343. [CrossRef]
6. Jiang, Z. Towards More Factual Large Language Models: Parametric and Non-Parametric Approaches. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2024.
7. Yang, L.; Yu, Z.; Zhang, T.; Cao, S.; Xu, M.; Zhang, W.; Gonzalez, J.E.; Cui, B. Buffer of thoughts: Thought-augmented reasoning with large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 113519–113544.
8. Sauvage, E. SynKGP: Knowledge Graph Population with Syntactic-LLM Hybridation for Question-Answering. In *European Conference on Information Retrieval*; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 212–219.
9. Aggarwal, V. Empowering Large Language Model Reasoning: Hybridizing Layered Retrieval Augmented Generation and Knowledge Graph Synthesis. *Int. J. High Sch. Res.* **2024**, *6*, 80–92. [CrossRef]
10. Zhang, Z.; Li, Y.; Le, N.H.L.; Wang, Z.; Ma, T.; Galassi, V.; Murugesan, K.; Moniz, N.; Geyer, W.; Chawla, N.V.; et al. NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning. *arXiv* **2024**, arXiv:2412.15547. [CrossRef]
11. Ma, C.; Chakrabarti, S.; Khan, A.; Molnár, B. Knowledge Graph-based Retrieval-Augmented Generation for Schema Matching. *arXiv* **2025**, arXiv:2501.08686.
12. Jiang, D.; Wang, H.; Li, T.; Gouda, M.A.; Zhou, B. Real-time tracker of chicken for poultry based on attention mechanism-enhanced YOLO-Chicken algorithm. *Comput. Electron. Agric.* **2025**, *237*, 110640. [CrossRef]
13. Mutabazi, E.; Ni, J.; Tang, G.; Cao, W. A review on medical textual question answering systems based on deep learning approaches. *Appl. Sci.* **2021**, *11*, 5456. [CrossRef]
14. Patil, R. and Gudivada, V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl. Sci.* **2024**, *14*, 2074. [CrossRef]
15. Lecu, A.; Groza, A.; Hawizy, L. Using LLMs and ontologies to extract causal relationships from medical abstracts. *Procedia Comput. Sci.* **2024**, *244*, 443–452. [CrossRef]
16. Jin, J.; Zhu, Y.; Dou, Z.; Dong, G.; Yang, X.; Zhang, C.; Zhao, T.; Yang, Z.; Wen, J.R. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In Proceedings of the Companion Proceedings of the ACM on Web Conference 2025, New York, NY, USA, 28 April–2 May 2025; pp. 737–740.
17. Wang, X.; Yang, Q.; Qiu, Y.; Liang, J.; He, Q.; Gu, Z.; Xiao, Y.; Wang, W. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv* **2023**, arXiv:2308.11761. [CrossRef]
18. Bussotti, J.F.; Ragazzi, L.; Frisoni, G.; Moro, G.; Papotti, P. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 12105–12122.
19. Yang, C.; Li, C.; Hu, X.; Yu, H.; Lu, J. Enhancing knowledge graph interactions: A comprehensive Text-to-Cypher pipeline with large language models. *Inf. Process. Manag.* **2026**, *63*, 104280. [CrossRef]
20. Longo, C.F.; Santamaria, D.F.; Mongiovì, M.; Bulla, L.; Sanfilippo, E.M. Leveraging knowledge graphs inference for semi-explainable systems based on large language models. In Proceedings of the Joint Ontology Workshops (JOWO)-Episode X: The Tukker Zomer of Ontology, and Satellite Events Co-Located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), Enschede, The Netherlands, 15–19 July 2024; Volume 3882.
21. Wang, Y.; Shen, X.; Huang, Z.; Niu, L.; Ou, S. cLegal-QA: A Chinese legal question answering with natural language generation methods. *Complex Intell. Syst.* **2025**, *11*, 77. [CrossRef]
22. Dehal, R.S.; Sharma, M.; Rajabi, E. Knowledge Graphs and Their Reciprocal Relationship with Large Language Models. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 38. [CrossRef]
23. Wang, J.; Cheng, Z.; Yao, Q.; Liu, L.; Xu, D.; Hu, G. Bioinformatics and biomedical informatics with ChatGPT: Year one review. *Quant. Biol.* **2024**, *12*, 345–359. [CrossRef] [PubMed]
24. Li, D.; Yang, S.; Tan, Z.; Baik, J.Y.; Yun, S.; Lee, J.; Chacko, A.; Hou, B.; Duong-Tran, D.; Ding, Y.; et al. DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature. *arXiv* **2024**, arXiv:2405.04819.

25. Li, Y.; Zhang, R.; Liu, J. An enhanced prompt-based LLM reasoning scheme via knowledge graph-integrated collaboration. In *International Conference on Artificial Neural Networks*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 251–265.
26. Sui, Y.; He, Y.; Liu, N.; He, X.; Wang, K.; Hooi, B. FiDeLiS: Faithful Reasoning in Large Language Model for Knowledge Graph Question Answering. *arXiv* **2024**, arXiv:2405.13873.
27. Sun, T.; Carr, J.; Kazakov, D. A Hybrid Question Answering Model with Ontological Integration for Environmental Information. In Proceedings of the DAO-XAI 2024: Workshop on Data Meets Applied Ontologies in Explainable AI, Santiago de Compostela, Spain, 19 October 2024.
28. Ji, Y.; Wu, K.; Li, J.; Chen, W.; Zhong, M.; Jia, X.; Zhang, M. Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 7598–7610.
29. Muqtadir, A.; Bilal, H.S.M.; Yousaf, A.; Ahmed, H.F.; Hussain, J. Mitigating Hallucinations Using Ensemble of Knowledge Graph and Vector Store in Large Language Models to Enhance Mental Health Support. *arXiv* **2024**, arXiv:2410.10853.
30. Liang, Y.; Tan, K.; Xie, T.; Tao, W.; Wang, S.; Lan, Y.; Qian, W. Aligning Large Language Models to a Domain-specific Graph Database for NL2GQL. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 1367–1377.
31. Nassiri, K.; Akhloufi, M. Transformer models used for text-based question answering systems. *Appl. Intell.* **2023**, *53*, 10602–10635. [CrossRef]
32. Badenes-Olmedo, C.; Corcho, O. MuHeQA: Zero-shot question answering over multiple and heterogeneous knowledge bases. *Semant. Web* **2024**, *15*, 1547–1561. [CrossRef]
33. Pusch, L.; Conrad, T.O. Combining LLMs and Knowledge Graphs to Reduce Hallucinations in Question Answering. *arXiv* **2024**, arXiv:2409.04181. [CrossRef]
34. Tian, Y.; Xia, F.; Song, Y. Two Heads are Better than One: Retrieval Augmented LLM for Question Answering with External Knowledge Attention. In Proceedings of the The Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
35. Wu, Y.; Huang, Y.; Hu, N.; Hua, Y.; Qi, G.; Chen, J.; Pan, J.Z. Cotkr: Chain-of-thought enhanced knowledge rewriting for complex knowledge graph question answering. *arXiv* **2024**, arXiv:2409.19753.
36. DeBellis, M.; Dutta, N.; Gino, J.; Balaji, A. Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation. *Appl. Ontol.* **2025**, 15705838241296446. [CrossRef]
37. Zhan, B.; Li, A.; Yang, X.; He, D.; Duan, Y.; Yan, S. RARoK: Retrieval-Augmented Reasoning on Knowledge for Medical Question Answering. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 2837–2843.
38. Doan, N.N.; Härmä, A.; Celebi, R.; Gottardo, V. A Hybrid Retrieval Approach for Advancing Retrieval-Augmented Generation Systems. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), Trento, Italy, 19–20 October 2024; pp. 397–409.
39. Xu, D.; Li, X.; Zhang, Z.; Lin, Z.; Zhu, Z.; Zheng, Z.; Wu, X.; Zhao, X.; Xu, T.; Chen, E. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. In Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 25570–25578.
40. Kim, S. MedBioLM: Optimizing Medical and Biological QA with Fine-Tuned Large Language Models and Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2502.03004.
41. Shi, Y.; Xu, S.; Yang, T.; Liu, Z.; Liu, T.; Li, X.; Liu, N. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. *AMIA Annu. Symp. Proc.* **2025**, *2024*, 1011.
42. Alonso, I.; Oronoz, M.; Agerri, R. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artif. Intell. Med.* **2024**, *155*, 102938. [CrossRef] [PubMed]
43. SNOMED CT. Available online: https://www.snomed.org/what-is-snomed-ct (accessed on 13 December 2023).
44. LOINC. Available online: https://loinc.org (accessed on 13 December 2023).
45. MeSH at NLM. Available online: https://meshb.nlm.nih.gov/search (accessed on 13 December 2023).
46. RxNorm. Available online: https://www.nlm.nih.gov/research/umls/rxnorm/index.html (accessed on 13 December 2023).
47. UMLS. Available online: https://www.nlm.nih.gov/research/umls/ (accessed on 13 December 2023).
48. Gene Ontology. Available online: http://geneontology.org (accessed on 13 December 2023).
49. PubMed Central (PMC). Available online: https://www.ncbi.nlm.nih.gov/pmc/ (accessed on 13 December 2023).
50. ClinicalTrials.gov. Available online: https://clinicaltrials.gov/ (accessed on 13 December 2023).
51. Radiopaedia. Available online: https://radiopaedia.org/ (accessed on 13 December 2023).
52. Open-i. Available online: https://openi.nlm.nih.gov/ (accessed on 13 December 2023).
53. MEDQA Dataset. Available online: https://huggingface.co/datasets/bigbio/mediqa_qa (accessed on 31 January 2024).
54. Available online: https://huggingface.co/docs/transformers/en/model_doc/gpt_neo (accessed on 12 November 2023).

55. Malin, B.; Kalganova, T.; Boulgouris, N. A review of faithfulness metrics for hallucination assessment in Large Language Models. *IEEE J. Sel. Top. Signal Process.* 2025; *early access*. [CrossRef]
56. Wood, M.C.; Forbes, A.A. 100% Hallucination Elimination Using Acurai. *arXiv* **2024**, arXiv:2412.05223. [CrossRef]
57. Geng, X.; Wang, H.; Wang, J.; Liu, W.; Li, R. Enhancing RAG with Active Learning on Conversation Records: Reject Incapables and Answer Capables. *arXiv* **2025**, arXiv:2502.09073.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.