Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Explainable Cardiovascular Question Answering with Controlled Knowledge-Graph Retrieval

**Mohammed Alshammari**

mohammed.k.alshammari@durham.ac.uk

Durham University

**Amir Atapour-Abarghouei**

Durham University

**Research Article**

**Keywords:**

**Posted Date:** November 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11402371/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Explainable Cardiovascular Question Answering with Controlled Knowledge-Graph Retrieval

Mohammed Alshammari[1*] and Amir Atapour-Abarghouei[1]

[1*]Department of Computer Science, Durham University, Stockton Road, Durham, DH1 3LE, United Kingdom .

*Corresponding author(s). E-mail(s): mohammed.k.alshammari@durham.ac.uk;
Contributing authors: amir.atapour-abarghouei@durham.ac.uk;

**Abstract**

Medical question-answering systems must provide accurate answers while maintaining traceability to clinical evidence. This work presents a cardiovascular question-answering framework combining query segmentation, explicit disease–aspect routing, controlled knowledge-graph retrieval, out-of-KG query rejection, source provenance, and route-level explainability. Clinical information from the NHS, British Heart Foundation, and MedlinePlus was organised into a knowledge graph with source-quality information and provenance. Routing combines lexical evidence with complementary pretrained semantic encoders, with T5-based reformulation applied when routing strength falls below a predefined threshold. Accepted disease–aspect routes retrieve stored clinical evidence rather than generating the final medical content. Causal Token Role Attribution (CTRA) explains how individual words in the accepted query formulation influence the disease–aspect routing decision. The framework was evaluated on a 577-question benchmark covering straightforward, long and multi-clause, spelling-error, MedQuAD, and out-of-KG questions. The proposed system achieved 97% disease Recall@1 and 95% aspect Recall@1 on long questions, obtained the highest BERTScore F1 across all evaluated in-graph categories, and returned no unsupported medical answer for any of the 100 evaluated out-of-KG questions. CTRA achieved higher explanation sufficiency and consistency than LIME while requiring substantially lower runtime.

**Keywords:** Medical question answering system, Knowledge graph, Explainable AI, Cardiovascular disease, Controlled knowledge retrieval

# 1 Introduction

Artificial intelligence (AI) is increasingly being used to improve access to medical information and support interaction with complex healthcare knowledge [1]. Pretrained language models have also been incorporated into medical question-answering systems to improve semantic interpretation and question–answer matching [2]. However, hallucination and factual inconsistency remain important concerns in generated responses [3, 4], while transparency and trust remain important considerations in healthcare AI [5]. These challenges are particularly important when medical information appears clinically plausible but is unsupported, outdated, or derived from sources of uncertain quality. Medical question-answering systems therefore require not only effective language understanding, but also grounded evidence, identifiable sources, provenance, and explanations of how the retrieved information was selected.

Knowledge graphs (KGs) provide a promising foundation because they represent biomedical information through explicit entities and relationships and support the integration of heterogeneous medical knowledge [6–8]. However, grounding an answer in a KG does not by itself address all dimensions of transparency. Provenance identifies where information originated, but does not indicate how the supporting source was assessed. Similarly, successful retrieval does not explain why a particular disease or clinical aspect was selected from the user's question. A gap therefore remains in making the path from a user's question to the retrieved evidence transparent enough to show why that evidence was selected and how its source was assessed.

To address this, we propose an explainable and source-quality-aware cardiovascular question-answering framework grounded in a curated cardiovascular knowledge graph. Cardiovascular medicine was selected as a focused test domain because it contains a broad range of conditions and recurring clinical information needs while still allowing the knowledge base to be curated from established public health sources. This provides a clinically coherent setting for evaluating disease–aspect routing and source-linked retrieval without making the underlying methodology specific to cardiovascular medicine. Clinical information is obtained from the National Health Service (NHS), British Heart Foundation (BHF), and MedlinePlus [9–11]. These clinical information sources are assessed using Currency, Relevance, Authority, Accuracy, and Purpose criteria [12, 13]. The resulting score represents source quality rather than model confidence and is retained together with provenance and supporting references.

The framework uses a hybrid disease–aspect routing strategy that combines explicit disease-name evidence, fuzzy lexical matching, and complementary semantic representations, whose contributions are adaptively weighted according to the query (Section 3.7). This enables the system to handle explicit disease mentions, lexical variation, misspellings, and semantically expressed questions. Once the disease and requested clinical aspect are accepted, they determine a controlled knowledge-graph retrieval route. The language-processing components therefore determine *where* to retrieve, while the knowledge graph determines *what* medical evidence is returned. The final medical content is retrieved from the knowledge graph rather than freely generated. Queries for which a sufficiently supported disease–aspect route cannot be established, including queries outside the supported knowledge-graph scope, are rejected rather than answered with unsupported content.

A central methodological contribution is Causal Token Role Attribution (CTRA), a routing-aligned explanation mechanism that applies controlled word-level interventions to the query formulation that produced the accepted disease–aspect route (Section 3.10). This formulation is the processed original query when it satisfies the routing threshold, or the selected T5 reformulation when recovery is required. CTRA identifies words that support the selected disease or clinical aspect, indicate competing interpretations, or alter the selected route when removed. Here, causal attribution refers specifically to the measured effect of controlled input interventions on routing behaviour, not to clinical or biological causality.

The main contributions of this work are as follows:

- A source-quality-aware cardiovascular KGQA framework that retains clinical provenance, supporting references, and structured source assessment within the knowledge graph (Sections 3.2 and 3.3).
- A controlled query-processing pipeline that separates multi-request questions into independently routable clinical requests while preserving the relevant disease context (Section 3.6).
- An adaptive disease–aspect routing mechanism that combines exact disease-name evidence, fuzzy lexical matching, and complementary semantic representations, with query-specific weighting of the available evidence (Section 3.7).
- A route-strength-based recovery mechanism that reformulates low-support queries and rejects them when a sufficiently supported disease–aspect route cannot be established to prevent unsupported retrieval (Section 3.8).
- A controlled knowledge-graph retrieval mechanism in which accepted disease–aspect routes select predefined Cypher queries, so that the returned medical content is retrieved directly from source-linked KG evidence rather than freely generated (Section 3.9).
- Causal Token Role Attribution (CTRA), a routing-aligned explanation method that measures how individual words in the accepted query formulation affect the accepted disease–aspect decision through controlled word-removal interventions (Section 3.10).

## 2 Related Work

Healthcare knowledge graphs have been used to integrate biomedical information and support retrieval, clinical knowledge integration, and question answering [6–8]. They have also been applied in specialised healthcare settings, including vaccine-safety information integration [14]. Early medical KGQA systems relied on structured graphs together with conventional NLP methods such as feature extraction, intent classification, entity recognition, and graph matching [15, 16]. These approaches demonstrated the value of structured medical knowledge, but their dependence on lexical and predefined processing rules can limit performance on misspellings and semantically varied questions.

Later approaches incorporated pretrained language models, semantic retrieval, and knowledge-graph information to improve question understanding and answer selection [2]. More recent work has combined structured knowledge-graph retrieval,

guideline-oriented retrieval-augmented generation, and LLM-based answer synthesis while exposing graph evidence and retrieved guideline passages to support answer traceability [17]. Other recent studies have explored LLMs with structured knowledge through logical-form generation [18], verifiable knowledge-graph reasoning paths [19], and LLM-assisted knowledge-graph construction and enrichment [20]. Similar KG–LLM integration has also been investigated in domain-specific question answering outside healthcare [21]. These studies demonstrate the growing integration of language models with structured knowledge; however, a remaining methodological concern is the preservation of all information requests in long or multi-clause questions when they are compressed into a single generated representation or synthesis step. In addition, generative reformulation or answer synthesis can make the final response differ from the wording of the retrieved evidence, which may reduce direct traceability to the original source content. These considerations motivate the present framework, which preserves disease–aspect requests separately and retrieves the final clinical evidence directly from the knowledge graph without generative rewriting.

Explainability remains an important issue. LIME explains predictions through input perturbation and a local surrogate model [22], while ERASER emphasises the importance of faithful explanations [23]. In healthcare NLP, interpretability remains a significant challenge [24]. KG-based reasoning paths can improve transparency [19], but they do not necessarily explain which words in the original query caused a particular disease or clinical aspect to be selected.

The main gap is therefore not only retrieving medical knowledge, but connecting query interpretation, controlled retrieval, provenance, source-quality assessment, and explanation within one traceable pipeline. The proposed framework addresses this by explicitly separating disease and clinical-aspect routing, retrieving evidence from a controlled knowledge-graph route, retaining source provenance and source-quality information, and using CTRA to explain how words in the accepted query formulation influence the routing decision.

# 3 Methods

## 3.1 Data Collection

Clinical text and disease-related images were collected through separate pipelines. Clinical content was obtained from the National Health Service (NHS), British Heart Foundation (BHF), and MedlinePlus [9–11]. The collected information was structured according to the clinical and provenance fields listed in Table 1.

Static webpages were extracted using Requests and BeautifulSoup [25], while dynamically rendered pages were processed using Selenium [26]. In accordance with the source-use policies, which require the clinical content to be retained without paraphrasing or summarisation, the collected text was preserved in its original wording. It was used only as knowledge-graph content and was not used to train the language models.

Disease-related images were collected separately from Radiopaedia [27] and Wikimedia Commons [28]. Radiopaedia was used only as an image source through a custom Python workflow with a Radiopaedia image-downloading tool [27]. For Wikimedia

Commons, images were collected for diseases for which a corresponding image could be identified [28]. Each image was linked to its corresponding disease, and the recorded URL, source, creator or contributor, and licence or usage information were retained as listed in Table 1.

The clinical text sources were assessed using the adapted CRAAP source-quality framework described in Section 3.2. The resulting score represents source quality, not model confidence or image quality. The dataset size and resulting knowledge-graph structure, including the final node and relationship counts, are reported in Section 3.3.

Figure 1 summarises the data-collection process.

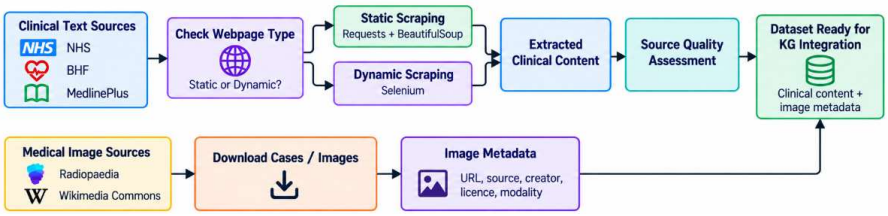


**Fig. 1** Data-collection pipeline for clinical text and disease-related images before knowledge-graph integration.

## 3.2 Source Quality Assessment

The clinical information sources were systematically assessed using an adapted CRAAP framework to evaluate their quality and suitability for inclusion in the knowledge graph. The assessment covered five dimensions: Currency, Relevance, Authority, Accuracy, and Purpose [12, 13]. These criteria were used specifically to assess the reliability and quality of each information source, independently of disease–aspect routing, route strength, or model confidence. The complete assessment criteria are provided in Supplementary File 1.

Each CRAAP dimension comprised five items scored from 0 to 2, giving a maximum of 10 points per dimension and an overall maximum score of 50:

$$S_{\text{CRAAP}} = C + R + A_u + A_c + P, \tag{1}$$

where $C$, $R$, $A_u$, $A_c$, and $P$ represent Currency, Relevance, Authority, Accuracy, and Purpose, respectively. The total score was used to classify the quality of each source as 45–50 = Excellent, 40–44 = Good, 35–39 = Usable with Caution, 30–34 = Poor, and $< 30$ = Unacceptable.

The CRAAP score and quality classification were represented through a Source Rating node linked to the corresponding Source node in the knowledge graph. This allows retrieved clinical evidence to include the quality assessment of its originating source. The CRAAP assessment represents source quality only; it does not affect disease–aspect routing, route strength, or retrieval ranking.

Currency, Relevance, and Authority were assessed using predefined reproducible criteria rather than subjective medical-expert judgement. Accuracy was independently

**Table 1** Clinical information and image metadata included in the cardiovascular dataset.

| Field | Description |
|---|---|
| **Clinical content** | |
| Disease Name | Cardiovascular disease name. |
| Overview | General disease information. |
| Symptoms | Associated signs and symptoms. |
| When to See a Doctor | Guidance on seeking medical attention. |
| Causes | Causes and contributing factors. |
| Complications | Associated complications. |
| Prevention | Prevention and risk-reduction information. |
| Diagnosis | Diagnostic information. |
| Treatment | Treatment and management information. |
| **Clinical provenance** | |
| Source | Clinical information provider. |
| Source URL | URL of the clinical webpage. |
| Date | Publication or update date provided by the source. |
| Source Quality Rating | CRAAP-based quality assessment of the associated clinical information source. |
| References | Supporting references provided by the source. |
| **Image provenance** | |
| Image URL | URL associated with the image. |
| Image Source | Source from which the image was obtained. |
| Image Creator / Contributor | Creator or contributor recorded by the image source. |
| Image Licence / Usage Information | Licence or usage information provided by the image source. |

assessed by three medical experts, while Purpose was assessed by both the first and second authors.

**Currency** was assessed using publication or last-review dates available in the dataset metadata. The scoring was applied consistently according to predefined age intervals: content published or updated within 0–5 years received 2 points for the publication-date item, content from 6–7 years received 1 point, and content older than 7 years received 0 points.

**Relevance** was assessed using predefined cardiovascular terminology derived from Medical Subject Headings (MeSH). Cardiovascular-related terms were checked across the available disease fields, including disease name, overview, symptoms, causes, diagnosis, treatment, complications, and prevention. This procedure was used to determine whether the source content was directly relevant to the cardiovascular scope of the knowledge graph.

**Authority** was controlled through source selection and the Authority checklist. Clinical content was restricted to established public medical organisations—NHS, MedlinePlus, and BHF—and the Authority criteria considered identification of the

responsible organisation, contact information, organisational affiliation, verifiable credentials, and subject expertise. This ensured that the Authority assessment was applied consistently to the institutional sources included in the knowledge graph.

**Accuracy** was independently assessed by three medical experts using the five Accuracy items. Each expert assigned a score out of 10, and the mean of the three assessments was used as $A_c$. The resulting Accuracy scores were 9.67 for NHS, 9.00 for MedlinePlus, and 9.00 for BHF.

**Purpose** was assessed independently by the first and second authors using five criteria covering clarity of purpose, objectivity, absence of inappropriate bias, factual presentation, and balanced communication. Each criterion was scored as 2 when satisfied and 0 when not satisfied, giving a maximum Purpose score of 10. Any uncertainty was discussed before assigning the final score.

Table 2 presents the final source-quality scores. All three clinical sources achieved an Excellent rating and were therefore retained in the knowledge graph.

**Table 2** CRAAP-based source-quality assessment.

| Source | Currency | Relevance | Authority | Accuracy | Purpose | Total | Rating |
|---|---|---|---|---|---|---|---|
| NHS | 10.00 | 10.00 | 10.00 | 9.67 | 10.00 | 49.67 | Excellent |
| MedlinePlus | 10.00 | 10.00 | 10.00 | 9.00 | 10.00 | 49.00 | Excellent |
| BHF | 9.00 | 10.00 | 10.00 | 9.00 | 10.00 | 48.00 | Excellent |

## 3.3 Knowledge Graph Construction

The curated cardiovascular dataset was transformed into a property graph implemented in Neo4j [29]. Each cardiovascular condition was represented as a *Disease* node. Its clinical information was represented using separate nodes for the eight supported aspects: overview, symptoms, when to see a doctor, causes, complications, prevention, diagnosis, and treatment. Each Disease node was connected to these clinical nodes through an explicit relationship corresponding to the information type, such as `HAS_SYMPTOMS`, `HAS_DIAGNOSIS`, or `HAS_TREATMENT`. This structure allows an accepted disease–aspect route to retrieve the corresponding clinical information directly from the graph.

Supporting provenance was represented through a separate *Metadata* node connected to each Disease. The Metadata node links the disease to its associated *Source*, *Reference*, and *Image* information. Each Source node is further linked to its *Source Rating*, which contains the CRAAP-based source-quality assessment described in Section 3.2. Source-quality ratings are retained as provenance information and do not influence disease–aspect routing or retrieval ranking.

Image provenance was represented explicitly. Each Image node contains links to the recorded image URL and source, together with creator or contributor and licence information provided in the source record.

Figure 2 illustrates the resulting graph structure.

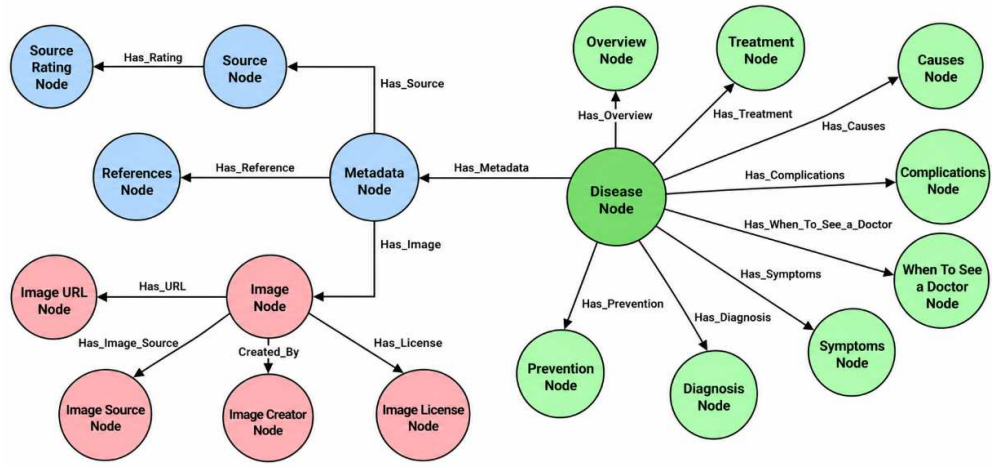


**Fig. 2** Structure of the cardiovascular knowledge graph.

The graph was populated automatically from the structured dataset using the Neo4j Python driver. `MERGE` operations were used to create or reuse nodes and relationships, reducing duplication during repeated ingestion. Reference and image records were deduplicated before being added to the graph.

The final knowledge graph contained 1,876 nodes and 3,707 relationships. This graph forms the retrieval layer of the question-answering system. Once a disease–aspect route is accepted, the corresponding clinical node is retrieved directly from the graph, while the associated Metadata node provides source provenance, source-quality information, references, and available image information.

## 3.4 Overall Question-Answering Framework

Following knowledge-graph construction, user questions are processed through a controlled question-routing and evidence-retrieval pipeline. Figure 3 provides an overview of the complete framework, including query preprocessing, controlled segmentation, disease and clinical-aspect routing, route-strength assessment, T5-based recovery for low-support queries, controlled knowledge-graph retrieval, and CTRA-based explanation. If a satisfactory route cannot be identified after recovery, the system returns a safe refusal. The individual components are described in the following subsections.

## 3.5 Query Preprocessing

Query preprocessing was used to reduce spelling and grammatical variations that could affect disease–aspect routing. Spelling and grammatical errors were corrected using GECToR [30] before the query was passed to the subsequent routing stages.

GECToR was selected because it applies targeted edits rather than rewriting the entire question, helping to preserve the original clinical meaning. Recognised cardiovascular disease names were protected during correction and restored before the corrected

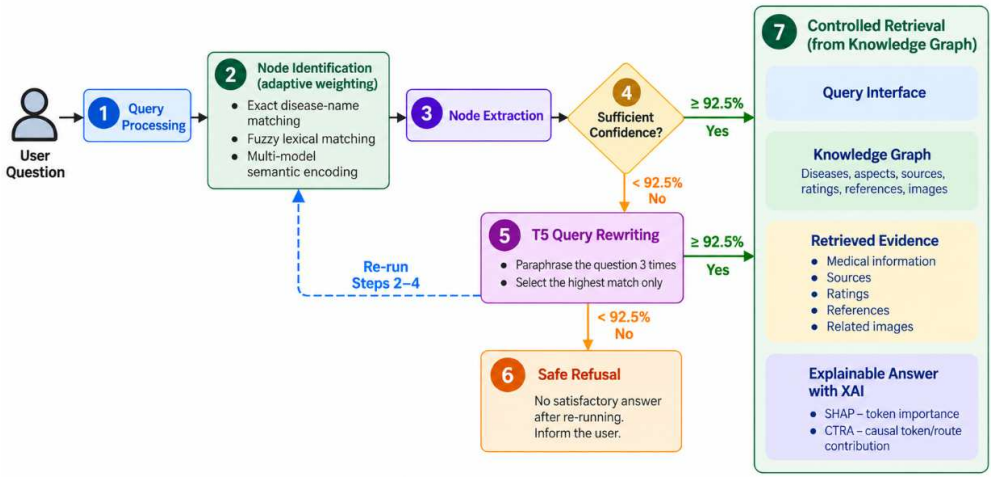


**Fig. 3** Overview of the proposed cardiovascular question-answering framework, from query processing and disease–aspect routing to route-strength-based recovery, controlled knowledge-graph retrieval, and routing-aligned explanation.

query was passed to controlled segmentation and disease–aspect routing. The preprocessing stage therefore modifies only the linguistic form of the query and does not determine the disease, clinical aspect, or medical answer.

## 3.6 Controlled Query Segmentation

Controlled query segmentation was used for questions containing more than one clinical request. Segmentation was applied to the corrected query $q^c$:

$$S(q^c) = \{s_1, s_2, \ldots, s_K\}, \tag{2}$$

where each $s_k$ represents an independently routable clinical request.

Candidate boundaries were identified at sentence boundaries and connectors such as commas, “and”, and “or”. A split was accepted only when each resulting segment contained a supported clinical-aspect intent, identified using the same predefined lexical aspect cues described in Section 3.7, and could be associated with a supported disease–aspect relation in the knowledge graph. Otherwise, the query remained unsplit.

When a segment expressed an aspect but did not repeat the disease name, the nearest explicitly mentioned disease in the surrounding query context was carried forward. If the segment explicitly mentioned a different disease, that disease was used instead. Full disease–aspect routing was then performed independently on each accepted segment, and duplicate routes were merged before retrieval.

This procedure preserves multiple valid clinical requests while avoiding unnecessary segmentation.

### 3.7 Hybrid Disease–Aspect Routing

Each processed query segment is routed independently to a disease and to the clinical aspect or aspects requested by the user. Disease and aspect routing are treated separately because identifying the medical condition and identifying the requested type of clinical information are distinct decisions. The resulting disease–aspect route determines the predefined relationship used during controlled knowledge-graph retrieval.

**Disease routing.** Disease routing combines exact disease-name matching, fuzzy lexical matching, and semantic matching. For a processed query segment $q$ and candidate disease $d$, exact lexical evidence is defined as

$$X(q,d) = \begin{cases} 1, & \text{if the complete normalised disease name occurs in } q, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Fuzzy lexical matching was implemented using RapidFuzz `fuzz.token_set_ratio`, with scores scaled to $[0,1]$:

$$S_{\text{lex}}(q,d) \in [0,1]. \tag{4}$$

Semantic matching used three complementary pretrained encoders: PubMedBERT/BiomedBERT (`microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext`) [31], BioBERT (`dmis-lab/biobert-base-cased-v1.1`) [32], and a Sentence Transformer (`sentence-transformers/all-mpnet-base-v2`) [33]. These pretrained checkpoints were used without task-specific fine-tuning. Disease candidates were represented by their normalised disease names, while each query segment was represented using its complete processed text.

For PubMedBERT/BiomedBERT and BioBERT, the corresponding pretrained tokeniser and transformer checkpoint were loaded directly. Inputs were tokenised with truncation enabled and a maximum sequence length of 512 tokens. The final hidden-state token representations were mean-pooled to obtain one fixed-length representation for each input. The Sentence Transformer was loaded using the `SentenceTransformer` implementation with the sentence-transformers/all-mpnet-base-v2 checkpoint, and its standard encoding procedure was used to obtain the corresponding fixed-length sentence representation. The same three pretrained checkpoints and encoding procedures were used throughout disease routing, semantic clinical-aspect routing, and subsequent rerouting operations.

Cosine similarity was calculated independently for each encoder between the query representation and every candidate disease representation. Let

$$s_e(q,d) \tag{5}$$

denote the semantic similarity assigned by encoder $e$ to candidate disease $d$.

Because the encoders can produce similarity scores on different numerical scales, the scores from each encoder were independently normalised across the complete disease candidate set:

$$\hat{s}_e(q,d) = \frac{s_e(q,d) - \min_{d' \in \mathcal{D}} s_e(q,d')}{\max_{d' \in \mathcal{D}} s_e(q,d') - \min_{d' \in \mathcal{D}} s_e(q,d') + \epsilon}, \tag{6}$$

where $\mathcal{D}$ is the supported disease candidate set and $\epsilon$ prevents division by zero.

The three semantic encoders begin with equal prior contributions:

$$p_e = \frac{1}{3}. \tag{7}$$

Equal priors provide a neutral starting point, but the final encoder contributions are adapted independently for each query according to how clearly each encoder separates its strongest disease candidate from its nearest alternative.

For encoder $e$, candidate separation is defined as

$$g_e(q) = \hat{s}_{e,(1)}(q) - \hat{s}_{e,(2)}(q), \tag{8}$$

where $\hat{s}_{e,(1)}(q)$ and $\hat{s}_{e,(2)}(q)$ are the highest and second-highest normalised disease-candidate scores. The separation is standardised and bounded as

$$z_e(q) = \frac{\max(0, g_e(q))}{\sigma_e(q) + \epsilon}, \tag{9}$$

$$\delta_e(q) = \tanh\left(z_e(q)\right), \tag{10}$$

where $\sigma_e(q)$ is the standard deviation of the encoder's normalised candidate scores. A larger $\delta_e(q)$ therefore indicates that the encoder more clearly separates its leading disease candidate from the alternatives.

The mean discrimination across the three encoders is

$$\bar{\delta}(q) = \frac{1}{3}\sum_{e=1}^{3} \delta_e(q). \tag{11}$$

The query-specific contribution of encoder $e$ is then calculated as

$$u_e(q) = p_e \exp\left[\beta\left(\delta_e(q) - \bar{\delta}(q)\right)\right], \tag{12}$$

followed by normalisation:

$$t_e(q) = \frac{u_e(q)}{\sum_{j=1}^{3} u_j(q)}. \tag{13}$$

The final encoder weight is

$$w_e(q) = (1 - \alpha)p_e + \alpha t_e(q). \tag{14}$$

The adaptive-routing parameters were fixed at $\alpha = 0.50$, $\beta = 1.50$, and $\epsilon = 10^{-9}$ before final evaluation and were kept unchanged across all experiments. These are fixed design parameters rather than learned parameters. The value $\alpha = 0.50$ balances the equal encoder prior with the query-specific adaptive contribution, while $\beta = 1.50$ controls the sensitivity of the weights to differences in encoder discrimination. The constant $\epsilon = 10^{-9}$ is used only for numerical stability.

The resulting encoder weights satisfy

$$\sum_{e=1}^{3} w_e(q) = 1. \tag{15}$$

If the three encoders provide equal candidate separation, their contributions remain 1/3. Otherwise, an encoder that distinguishes its leading candidate more clearly receives a relatively larger contribution for that query.

The combined semantic disease score is

$$S_{\text{sem}}(q,d) = \sum_{e=1}^{3} w_e(q)\hat{s}_e(q,d). \tag{16}$$

**Fusion of lexical and semantic disease evidence.** The combined semantic score is subsequently integrated with exact and fuzzy lexical evidence:

$$H_X(q,d) = X(q,d), \qquad H_L(q,d) = S_{\text{lex}}(q,d), \qquad H_S(q,d) = S_{\text{sem}}(q,d). \tag{17}$$

The fuzzy lexical and semantic evidence sources are always active. Exact lexical evidence is included only when a complete supported disease name is explicitly identified in the query segment.

For each active evidence source $k$, candidate separation is defined as

$$g_k^H(q) = H_{k,(1)}(q) - H_{k,(2)}(q), \tag{18}$$

where $H_{k,(1)}(q)$ and $H_{k,(2)}(q)$ are the highest and second-highest disease-candidate scores produced by that evidence source.

The separation is standardised and bounded as

$$\delta_k^H(q) = \tanh\left(\frac{\max\left(0, g_k^H(q)\right)}{\sigma_k^H(q) + \epsilon}\right), \tag{19}$$

where $\sigma_k^H(q)$ is the standard deviation of the disease-candidate scores produced by evidence source $k$.

The contribution of each active evidence source is

$$\omega_k(q) = \frac{\delta_k^H(q)}{\sum_j \delta_j^H(q)}, \tag{20}$$

when the denominator is greater than zero. If all active evidence sources provide zero discrimination, equal contributions are used.

The final hybrid disease-routing score is

$$S_{\text{hybrid}}(q,d) = \sum_k \omega_k(q) H_k(q,d), \tag{21}$$

and the highest-scoring disease is

$$d^* = \arg\max_{d \in \mathcal{D}} S_{\text{hybrid}}(q,d). \tag{22}$$

The two adaptive operations serve different purposes. Equation 14 adapts the contributions of PubMedBERT/BiomedBERT, BioBERT, and the Sentence Transformer within the semantic score. Equation 21 then adapts the contributions of semantic,

fuzzy lexical, and available exact lexical evidence. Selection of $d^*$ does not itself accept the route; route strength is evaluated separately in Section 3.8.

**Clinical-aspect routing.** The knowledge graph supports eight clinical aspects: overview, symptoms, when to see a doctor, causes, complications, prevention, diagnosis, and treatment.

Aspect routing first checks predefined lexical cues and their lemmatised forms. When one or more explicit supported aspect cues are identified, those aspects are retained directly. Semantic matching is used when lexical evidence is insufficient to identify the requested aspect. For evaluation, explicitly identified aspects are placed ahead of aspects identified only through semantic matching, while the remaining aspects are ordered by their semantic scores.

Semantic aspect routing uses the same three pretrained encoders and encoding procedures described for disease routing. Each query segment is compared with representations of the eight predefined aspect labels, and the resulting semantic scores are used to rank the aspect candidates.

Let

$$s_e^A(q, a) \tag{23}$$

denote the semantic similarity assigned by encoder $e$ to candidate aspect $a$. For each encoder, the scores are independently normalised across the supported aspect candidate set:

$$\hat{s}_e^A(q, a) = \frac{s_e^A(q, a) - \min_{a' \in \mathcal{A}} s_e^A(q, a')}{\max_{a' \in \mathcal{A}} s_e^A(q, a') - \min_{a' \in \mathcal{A}} s_e^A(q, a') + \epsilon}, \tag{24}$$

where $\mathcal{A}$ contains the eight supported clinical aspects.

The three encoders again begin with equal priors,

$$p_e^A = \frac{1}{3}, \tag{25}$$

and the same candidate-separation and adaptive-weighting procedure used for disease routing is applied independently to the aspect candidates. This produces query-specific aspect weights $w_e^A(q)$ satisfying

$$\sum_{e=1}^{3} w_e^A(q) = 1. \tag{26}$$

The combined semantic aspect score is

$$S_A(q, a) = \sum_{e=1}^{3} w_e^A(q) \hat{s}_e^A(q, a). \tag{27}$$

When semantic fallback is required, the selected aspect is

$$a^* = \arg\max_{a \in \mathcal{A}} S_A(q, a). \tag{28}$$

Semantic fallback selects the single highest-scoring aspect for that query segment. Multiple aspects are retained only when they are explicitly identified through

lexical cues or when controlled segmentation separates the question into multiple independently routable requests.

Explicit lexical aspect evidence is therefore retained rather than replaced by semantic inference. If several supported aspects are explicitly identified, all are preserved and paired with the corresponding selected disease.

Overall, disease and clinical-aspect routing begin with equal semantic-encoder priors but adapt their final contributions independently for each query. Disease routing additionally combines semantic evidence with fuzzy and exact lexical evidence. The selected disease and clinical aspect form a candidate route whose support is subsequently evaluated using the route-strength and recovery procedure described in Section 3.8.

## 3.8 Bridging the Linguistic Gap Between Natural Language Input and Medical Knowledge Graphs

Informal, indirect, or semantically ambiguous questions can produce weakly supported disease–aspect routes. Before knowledge-graph retrieval, the system therefore evaluates the strength of the route produced by the hybrid router in Section 3.7. The contributions of PubMedBERT, BioBERT, and the Sentence Transformer remain query-adaptive as defined in that section; no additional encoder weighting is introduced during recovery.

For the selected disease $d^*$, the routing margin is

$$M_D(q) = S_{\text{hybrid}}(q, d^*) - \max_{d \neq d^*} S_{\text{hybrid}}(q, d), \tag{29}$$

and the corresponding disease support is

$$C_D = \tanh\left(\frac{\max(0, M_D(q))}{\sigma_D(q) + \epsilon}\right), \tag{30}$$

where $\sigma_D(q)$ is the standard deviation of the disease-candidate scores and $\epsilon$ prevents division by zero. A larger margin relative to the variation among candidate scores therefore indicates stronger support for the selected disease.

When semantic aspect routing is used, aspect support is calculated in the same manner:

$$M_A(q) = S_A(q, a^*) - \max_{a \neq a^*} S_A(q, a), \tag{31}$$

$$C_A = \tanh\left(\frac{\max(0, M_A(q))}{\sigma_A(q) + \epsilon}\right), \tag{32}$$

where $S_A(q, a)$ is the adaptive semantic aspect score defined in Section 3.7, and $\sigma_A(q)$ is the standard deviation of the aspect-candidate scores. When an explicit lexical cue directly identifies the requested aspect, $C_A = 1$ because the aspect is determined directly rather than through semantic fallback.

Disease and aspect support contribute equally to the overall route-strength score:

$$C_R = \frac{C_D + C_A}{2}. \tag{33}$$

The value $C_R$ measures the relative support for the selected disease–aspect route. It is used only to determine whether the route can proceed to retrieval and should not be interpreted as a probability of clinical correctness, model confidence, or source-quality score. Because $C_R$ is based on the separation between the selected candidate and the available alternatives, it evaluates how strongly the selected route is preferred within the supported candidate space. An unsupported condition may also produce a strong relative separation if it is substantially closer to one supported disease than to the remaining candidates. Accordingly, the threshold is used to control whether a route has sufficient support to proceed to retrieval, rather than to make an absolute determination of whether a condition is inside or outside the knowledge graph.

**Threshold selection.** The recovery threshold was selected using a separate held-out development set of 100 questions that did not overlap with the final evaluation benchmark. Thresholds of 97.5%, 95%, 92.5%, 90%, 87.5%, and 85% were evaluated. For each threshold, routing performance was calculated as the mean of disease-routing and clinical-aspect routing accuracy after applying the corresponding threshold-triggered recovery procedure.

As shown in Figure 4, routing performance remained at 100% for thresholds from 97.5% down to 92.5%. Below 92.5%, performance decreased to 90.0% at a 90% threshold, 87.5% at 87.5%, and 82.5% at 85%. Lowering the threshold allowed weaker initial routes to proceed without recovery, reducing routing performance on the development set. Therefore, 92.5% ($C_R = 0.925$) was selected as the lowest tested threshold that maintained the maximum observed performance. This value was fixed before evaluation on the final benchmark.

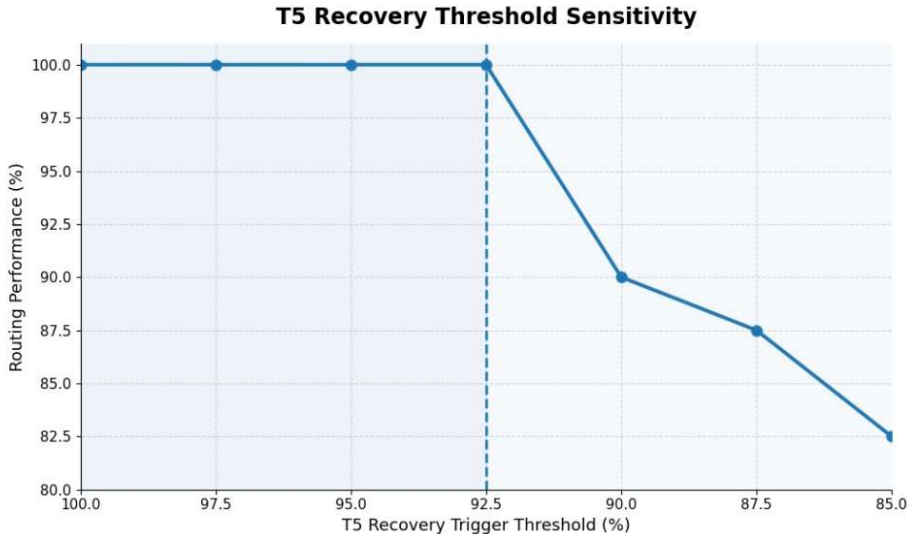


**Fig. 4** T5 recovery-threshold sensitivity on the held-out 100-question development set. Routing performance remained at 100% from thresholds 97.5% to 92.5% and decreased below 92.5%.

A route is accepted directly when

$$C_R \geq 0.925. \tag{34}$$

Accepted routes proceed to controlled knowledge-graph retrieval. When $C_R < 0.925$, the initial route is withheld and the query is passed to T5 [34] for linguistic reformulation. T5 is used only as a recovery mechanism. It does not independently select the disease or clinical aspect, retrieve medical evidence, generate a Cypher query, or produce the final medical answer.

The pretrained `ramsrigouthamg/t5_paraphraser` model was used without task-specific fine-tuning. Input sequences were truncated to a maximum of 64 tokens, and generated sequences were limited to 64 tokens. Generation used five beams with sampling enabled, temperature 0.9, top-$k = 50$, top-$p = 0.95$, three returned sequences, and early stopping. Because sampling introduces stochasticity, a fixed random seed of 42 was used for all T5 generation during threshold development, final benchmark evaluation, and explanation-consistency evaluation. The remaining generation settings were kept unchanged across all evaluated queries.

For each low-support query segment, three reformulations were generated. Each reformulation was independently processed through the same preprocessing and disease–aspect routing procedure. The query-adaptive encoder contributions and route-strength score were recalculated for every reformulation. The reformulation with the highest $C_R$ was retained. Three reformulations were fixed during development because additional candidates mainly produced repeated or near-duplicate formulations without improving routing.

If the strongest reformulated route reaches the same acceptance criterion, $C_R \geq 0.925$, it proceeds to controlled knowledge-graph retrieval. Otherwise, retrieval is stopped and the predefined safe-refusal response is returned:

> "Warning: We don't have a satisfactory answer for this question despite several paraphrasing attempts."

For multi-request questions, route strength is evaluated independently for each disease–aspect route produced by controlled segmentation. Any route below the 92.5% threshold must therefore pass through the same recovery procedure before retrieval. T5 can modify only the linguistic formulation of the query; it cannot bypass the routing threshold or independently determine or generate the medical content returned to the user.

## 3.9 Controlled Knowledge Graph Retrieval

Once the disease $d^*$ and clinical aspect $a^*$ are accepted by the routing stage, they determine the corresponding knowledge-graph retrieval route. Each supported clinical aspect is mapped to a predefined KG relationship, and the accepted disease and aspect are passed as parameters to the appropriate Cypher query.

All clinical Cypher queries are manually predefined. They are not generated by an LLM or by LlamaIndex. This keeps the retrieval process constrained to the disease–aspect relationships defined in the knowledge-graph schema and prevents unrestricted query generation.

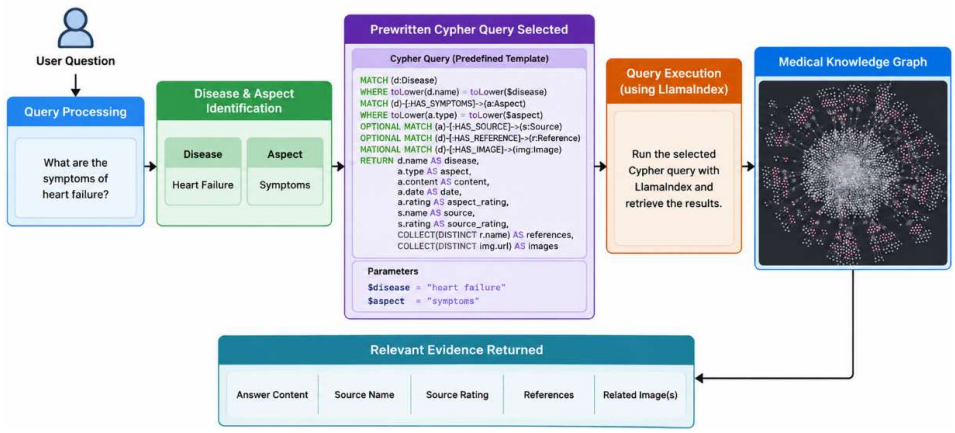


**Fig. 5** Controlled knowledge-graph retrieval. The identified disease and clinical aspect select a predefined Cypher query, which retrieves the corresponding clinical evidence and supporting metadata from the knowledge graph.

LlamaIndex [35] is used as a supporting integration layer for the property-graph and embedding infrastructure within the framework. It does not determine the disease or clinical aspect, select or generate the clinical Cypher query, or execute the final clinical retrieval. The accepted disease–aspect route selects the predefined query, which is executed against Neo4j through the database interface, as illustrated in Figure 5.

The accepted disease–aspect route retrieves the stored clinical content from the corresponding clinical node. Supporting provenance is retrieved through the Metadata node associated with the selected Disease. The Metadata node links to the corresponding Source, Reference, and Image information, while each Source is linked to its CRAAP-based Source Rating. Source quality does not participate in disease–aspect routing and is not used to rank or re-rank the retrieved clinical content.

The retrieved output therefore contains the clinical content associated with the accepted disease–aspect route together with the available source information, source-quality rating, references, and related images. The final medical content is retrieved directly from the knowledge graph rather than generated or summarised by a language model.

## 3.10 Causal Token Role Attribution

After a disease–aspect route is accepted, the proposed Causal Token Role Attribution (CTRA) method explains which words in the accepted query formulation influenced that routing decision. CTRA operates directly on the deployed disease–aspect router rather than fitting a separate surrogate model. Each analysed non-punctuation word is removed in turn, and the disease–aspect routing procedure is rerun on the perturbed formulation. T5 recovery and knowledge-graph answer retrieval are not repeated during this intervention because CTRA explains the routing decision that has already been accepted.

**Analysed query and route.** CTRA analyses the query formulation that produced the accepted route, referred to hereafter as the *accepted query formulation*. If

the processed original query satisfies the routing threshold, that formulation is analysed directly. If T5-based recovery is required, CTRA instead analyses the selected T5 reformulation that satisfies the routing criterion and produces the accepted route. CTRA therefore explains the routing decision from the formulation on which that decision was actually accepted, which is not necessarily the wording of the original user question.

For a single query segment, the accepted route is

$$R(q) = (d^*, a^*), \tag{35}$$

where $d^*$ and $a^*$ are the selected disease and clinical aspect for the accepted query formulation. Multi-request questions are analysed internally as separate segments. If a question produces several accepted disease–aspect routes, CTRA is applied independently to each accepted route and the resulting token attributions are subsequently aligned back to the complete accepted query formulation for presentation. This prevents one valid requested aspect from being incorrectly treated as a competitor to another valid aspect.

**Word-removal intervention.** For each non-punctuation word $w_i$, CTRA constructs $q_{-i}$ by removing that word and reruns the same disease–aspect routing procedure. The selected labels, their support scores, and their margins over competing candidates are then compared with those obtained from the complete accepted query formulation.

**Continuous routing-support profiles.** CTRA measures these changes using continuous candidate-score profiles obtained directly from the deployed router. Let $\widetilde{S}_D(q, d)$ and $\widetilde{S}_A(q, a)$ denote the routing-aligned continuous support profiles for disease and clinical-aspect candidates, respectively. These profiles are not additional prediction models or independently learned explanation scores; they are constructed from the same scores used by the deployed routing mechanism.

For disease routing, let $\mathcal{K}_D(q)$ denote the evidence sources that were active when the accepted query $q$ was routed. Fuzzy lexical and semantic evidence are always active, while exact lexical evidence is included when a complete supported disease name is identified. The disease support profile is therefore the hybrid disease-routing score defined in Section 3.7:

$$\widetilde{S}_D(q, d) = \sum_{k \in \mathcal{K}_D(q)} \omega_k(q) H_k(q, d), \tag{36}$$

where $H_k(q, d)$ is the candidate score produced by evidence source $k$ and $\omega_k(q)$ is its query-adaptive fusion weight. Thus, $\widetilde{S}_D(q, d)$ corresponds directly to the continuous score used to rank disease candidates.

During the intervention on $q_{-i}$, the set of evidence sources active for the accepted formulation, $\mathcal{K}_D(q)$, is held fixed so that CTRA continues to explain the routing mechanism that produced the original decision. The individual evidence scores, candidate separations, and fusion weights are then recalculated from the perturbed formulation:

$$\widetilde{S}_D(q_{-i}, d) = \sum_{k \in \mathcal{K}_D(q)} \omega_k(q_{-i}) H_k(q_{-i}, d). \tag{37}$$

Consequently, removing a word can weaken or strengthen exact, fuzzy, or semantic support and can change their adaptive contributions, while the comparison remains aligned with the evidence sources used for the accepted route.

For clinical-aspect routing, CTRA preserves the routing branch that produced the accepted aspect. If the aspect was selected through explicit lexical cues, let $h_a(q)$ denote the number of matched cues for aspect $a$. The continuous lexical aspect profile is

$$\widetilde{S}_A(q,a) = \frac{h_a(q)}{\max_{a' \in \mathcal{A}} h_{a'}(q)}, \tag{38}$$

where $\mathcal{A}$ is the set of supported clinical aspects. If no lexical cue remains after a word-removal intervention, the lexical profile contains zero support for all aspects.

If the accepted aspect was instead selected through semantic fallback, the support profile is the adaptive semantic aspect score defined in Section 3.7:

$$\widetilde{S}_A(q,a) = S_A(q,a). \tag{39}$$

The same original aspect branch—lexical or semantic—is retained when calculating the support profile for $q_{-i}$, and its candidate scores are recomputed after each word removal. This prevents CTRA from attributing a lexical decision to the semantic branch, or vice versa. Route-change detection itself is based on rerunning the deployed router normally, allowing CTRA to record whether the actual selected disease or aspect changes after the intervention.

For the selected disease,

$$S_D^*(q) = \widetilde{S}_D(q,d^*), \tag{40}$$

with margin

$$M_D(q) = \widetilde{S}_D(q,d^*) - \max_{d \neq d^*} \widetilde{S}_D(q,d). \tag{41}$$

The positive disease effect of removing word $w_i$ is

$$P_i^D = \frac{1}{2}\max\left(0, S_D^*(q) - S_D^*(q_{-i})\right) + \frac{1}{2}\max\left(0, M_D(q) - M_D(q_{-i})\right). \tag{42}$$

The opposite-direction effect is

$$N_i^D = \frac{1}{2}\max\left(0, S_D^*(q_{-i}) - S_D^*(q)\right) + \frac{1}{2}\max\left(0, M_D(q_{-i}) - M_D(q)\right). \tag{43}$$

Thus, $P_i^D > 0$ means that removing the word weakens support for the selected disease, whereas $N_i^D > 0$ means that removing it strengthens the selected disease route.

The same intervention is applied to the selected clinical aspect:

$$S_A^*(q) = \widetilde{S}_A(q,a^*), \tag{44}$$

$$M_A(q) = \widetilde{S}_A(q,a^*) - \max_{a \neq a^*} \widetilde{S}_A(q,a), \tag{45}$$

$$P_i^A = \frac{1}{2}\max\left(0, S_A^*(q) - S_A^*(q_{-i})\right) + \frac{1}{2}\max\left(0, M_A(q) - M_A(q_{-i})\right), \tag{46}$$

and

$$N_i^A = \frac{1}{2}\max\left(0, S_A^*(q_{-i}) - S_A^*(q)\right) + \frac{1}{2}\max\left(0, M_A(q_{-i}) - M_A(q)\right). \tag{47}$$

The signed routing effect of word $w_i$ is

$$I_i = \left(P_i^D + P_i^A\right) - \left(N_i^D + N_i^A\right). \tag{48}$$

A positive $I_i$ indicates that the word supports the accepted route because removing it weakens that route. A negative $I_i$ indicates movement in the opposite direction: removing the word strengthens the accepted route. Retaining the sign allows CTRA to distinguish supporting and opposing effects rather than discarding one direction of change.

CTRA also records whether removal changes the selected route:

$$B_i = \mathbb{I}\left[R(q_{-i}) \neq R(q)\right]. \tag{49}$$

A word with $B_i = 1$ is considered *load-bearing*, because removing it changes the selected disease, clinical aspect, or both.

Repeated occurrences of the same selected disease expression or explicit aspect cue are handled jointly. CTRA removes the repeated evidence as a group to measure its overall effect and distributes that effect across its occurrences using their individual deletion sensitivities. This prevents redundant wording from receiving artificially low attribution simply because another occurrence remains after one word is removed.

**Word roles.** The measured intervention determines the magnitude of a word's influence, while the routing structure determines its explanatory role. CTRA distinguishes:

- **Disease-condition evidence**: words belonging to the selected disease expression;
- **Clinical-aspect evidence**: words explicitly supporting the selected clinical aspect;
- **Competing-route evidence**: explicit words indicating a different disease or clinical aspect; and
- **Contextual information**: words that are not direct disease, aspect, or competing-route cues.

A word can support both the selected disease and the selected clinical aspect when its removal reduces support for both routing components. Importantly, a negative intervention value does not by itself identify competing-route evidence; a word is labelled as competing-route evidence only when it explicitly refers to a different disease or clinical aspect.

Disease evidence is displayed in green, aspect evidence in blue, competing-route evidence in red or pink, and contextual information in grey. Words modified during preprocessing retain their routing role and receive an additional orange correction marker. Punctuation is displayed for readability but excluded from attribution scoring.

For presentation, the magnitude of each signed intervention effect is converted into a relative impact weight:

$$W_i = \begin{cases} 100\dfrac{|I_i|}{\sum_{j\in\mathcal{W}} |I_j|}, & \text{if } \sum_{j\in\mathcal{W}} |I_j| > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{50}$$

where $\mathcal{W}$ contains all analysed non-punctuation words. For multi-segment questions, a common denominator is used after the segment-level results are aligned back to the complete accepted query formulation. The displayed impact weights therefore sum to 100% when measurable routing impact is present. These values represent the relative distribution of routing impact within the accepted query formulation and are not probabilities, confidence scores, or measures of clinical correctness.

**Explanation evaluation.** CTRA was evaluated using explanation sufficiency, explanation consistency, explanation length, and runtime. The same evaluation metrics and route-preservation criteria were applied to CTRA and LIME so that both methods were assessed against the same accepted disease–aspect routing decisions.

**LIME interface to the disease–aspect router.** LIME [22] requires a classifier-style numerical output for each perturbed input. To apply LIME to the same disease–aspect routing decision explained by CTRA, the deployed router was wrapped as a two-class route-preservation function. For the accepted target route $R(q) = (d^*, a^*)$, each LIME perturbation $q'$ was rerouted and classified according to

$$Z(q') = \mathbb{I}\left[R(q') = (d^*, a^*)\right]. \tag{51}$$

The two output classes were *route changed* and *route preserved*. When the perturbed formulation preserved both the selected disease and clinical aspect, the wrapper returned

$$[0.01,\ 0.99], \tag{52}$$

corresponding to *route changed* and *route preserved*, respectively. When either component of the target route changed, the wrapper returned

$$[0.99,\ 0.01]. \tag{53}$$

These values are interface scores used to provide LIME with the binary routing outcome required by its local surrogate model; they are not calibrated probabilities, route-strength scores, model confidence values, or estimates of clinical correctness. LIME therefore explains preservation of the same accepted disease–aspect decision as CTRA, but approximates that decision from perturbed samples using a locally fitted surrogate model.

T5 recovery and knowledge-graph retrieval were not invoked for the LIME perturbations. LIME was configured with 300 perturbation samples per query and a fixed random seed of 42. The standard `LimeTextExplainer` settings were retained for parameters not explicitly specified. Up to 12 query features were exposed to the local surrogate model. The *route preserved* class was explained, and positively weighted features were ranked by the magnitude of their LIME weights. Up to the five highest-ranked positive words were retained as the LIME explanation; if fewer than five positively weighted words were available, all available positive words were retained.

For quantitative evaluation of CTRA, words with a positive signed routing effect ($I_i > 0$) were treated as supporting the accepted disease–aspect route. These words were ranked in descending order of their relative impact weight $W_i$, and up to the five highest-ranked words were retained as the explanation set $\mathcal{E}(q)$. If fewer than five words had a positive routing effect, all available positive words were retained. Words with zero or negative routing effects were not included in $\mathcal{E}(q)$, although their signed effects and explanatory roles remained available in the complete CTRA visualisation. Ties were resolved according to the word order in the accepted query formulation. For multi-request questions, segment-level CTRA results were first aligned to the complete accepted query formulation and the same five-word selection rule was then applied.

For evaluation, let $\mathcal{E}(q)$ denote the explanation words selected by the evaluated explanation method for query $q$, and let $q_{\mathcal{E}}$ denote the reduced query containing only those words. Explanation sufficiency is defined as

$$\mathrm{Suff}(q) = \mathbb{I}\left[R(q_{\mathcal{E}}) = R(q)\right]. \tag{54}$$

Sufficiency therefore equals 1 when the selected explanation words alone reproduce the accepted disease–aspect route and 0 otherwise. For $N$ evaluated questions, the reported sufficiency percentage is

$$\text{Sufficiency} = \frac{100}{N} \sum_{n=1}^{N} \mathrm{Suff}(q_n). \tag{55}$$

Explanation consistency evaluates whether alternative formulations that preserve the same accepted disease–aspect decision produce similar explanatory concepts. For each in-KG evaluation question, three alternative formulations were generated using the same T5 paraphraser and generation settings described in Section 3.8. Each alternative was independently rerouted through the disease–aspect routing procedure. An alternative was retained only when it satisfied the predefined routing threshold of 0.925 and produced the same accepted disease–aspect route as the reference formulation. For multi-request questions, the complete set of accepted disease–aspect routes had to be preserved. Route-changing, below-threshold, and exact duplicate alternatives were excluded.

The reference route was the route accepted by the deployed system: the processed original formulation when it satisfied the routing threshold, or the selected T5 reformulation when recovery was required. A question contributed to the consistency calculation only when at least two distinct route-preserving alternatives remained. Let $N_{\mathrm{cons}}$ denote the number of questions satisfying this requirement. Questions with fewer than two retained alternatives were excluded because pairwise similarity cannot be calculated. The value of $N_{\mathrm{cons}}$ is reported separately for each question category in the explainability results table.

Before comparison, explanation words were canonicalised by lemmatisation and by mapping lexical variants representing the same clinical aspect to a common concept. For two retained route-preserving formulations $q_r$ and $q_s$, let $\mathcal{C}_r$ and $\mathcal{C}_s$ denote their canonical explanation-concept sets. Pairwise consistency is calculated using Jaccard similarity:

$$J(q_r, q_s) = \frac{|\mathcal{C}_r \cap \mathcal{C}_s|}{|\mathcal{C}_r \cup \mathcal{C}_s|}. \tag{56}$$

When three route-preserving alternatives were retained, Jaccard similarity was calculated for all three pairwise combinations and averaged. When two alternatives were retained, the single available pairwise comparison was used. The reported consistency percentage is the mean of the question-level consistency values across the $N_{\text{cons}}$ eligible questions, multiplied by 100. A higher value indicates that route-preserving alternative formulations produce more similar explanatory concepts.

Explanation length is the number of unique explanation words selected for each query, averaged across the evaluation set. Runtime is the average time required to generate an explanation for one query. All CTRA and LIME runtime measurements were obtained on the same workstation equipped with an NVIDIA GeForce RTX 4070 GPU with 12 GB of VRAM. T5 alternative generation, rerouting of the alternative formulations, and pairwise consistency calculations were excluded from the reported explanation runtime because these operations form part of the evaluation procedure rather than generation of a single explanation.

The term *causal* refers only to the controlled intervention performed by CTRA: a word is removed, the disease–aspect router is rerun, and the resulting change in routing behaviour is measured. It does not imply biological, diagnostic, or clinical causality.

CTRA therefore provides a route-specific explanation that identifies both the role and relative impact of the words contributing to an accepted disease–aspect decision. A sample CTRA output can be found in the *Supplementary information*, File 2.

## 3.11 Evaluation Metrics and Performance Measures

The framework was evaluated using 577 questions covering straightforward, long and multi-clause, spelling-error, MedQuAD, and out-of-KG conditions.

**Benchmark construction.** Straightforward questions involved one supported cardiovascular disease and one clinical aspect. Long and multi-clause questions used more complex formulations and could contain multiple disease–aspect requests. Spelling-error questions contained misspelled disease expressions while preserving the intended disease and aspect. MedQuAD provided an external set of real-user medical questions [36]; only questions whose disease was represented in the knowledge graph and whose requested information matched one of the eight supported clinical aspects were selected. Out-of-KG questions concerned conditions absent from the knowledge graph, including cases containing terminology that overlapped with supported conditions. All in-KG questions were annotated with their expected disease and clinical aspect, with all valid targets retained for multi-request questions. The manually constructed questions and their annotations were produced by the authors and subsequently reviewed by the same three medical experts who assessed the CRAAP Accuracy component. The manually constructed benchmark and the selected MedQuAD evaluation questions are provided in the GitHub repository. A separate 100-question development set was used only for routing-threshold selection and did not overlap with the final benchmark.

**Baseline systems.** Five published approaches were reimplemented for comparison: KG-based medical QAS [15], medical KBQA [16], QASTRisk [21], cMedQA-based medical QA [2], and Hybrid KG–LLM QA [17]. Each used the same supported disease and clinical-aspect candidate space as the proposed system. GPT-4o, GPT-5.4, and MedGemma-27B-IT were also evaluated as zero-shot baselines without task-specific training. Their prompt required up to five ranked disease and clinical-aspect candidates for supported questions and no answer for out-of-KG questions. The complete prompt and model settings are provided in the accompanying GitHub repository.

**Routing metrics.** Disease and clinical-aspect routing were evaluated separately using Recall@1, Recall@5, mean reciprocal rank (MRR), and nDCG@5. Recall@1 measures whether the correct candidate is ranked first, while Recall@5 measures whether it appears within the top five [37, 38]. MRR rewards higher placement of the correct candidate [38], and nDCG@5 evaluates the quality of the top-five ranking [39]. For multi-request questions, each annotated disease–aspect pair was evaluated separately. Routes remaining below the acceptance threshold after T5 recovery were treated as incorrect.

**Answer-retrieval fidelity.** BERTScore Precision, Recall, and F1 [40] compared each returned answer with the corresponding reference clinical content stored in the knowledge graph. For multi-request questions, each disease–aspect route was evaluated against its corresponding KG reference content. BERTScore is therefore used here to measure retrieval fidelity and semantic agreement with the stored KG reference content, rather than independent clinical correctness.

**Explainability.** CTRA was compared with LIME [22] using sufficiency, consistency, explanation length, and runtime. For consistency, only distinct T5-generated alternatives preserving the accepted disease–aspect route were retained. The complete procedure is described in Section 3.10.

**Out-of-KG evaluation.** Out-of-KG performance measured whether a system returned medical information for a condition absent from the knowledge graph. Responses were classified as *answered* when unsupported medical content was returned and as *refused/no answer* otherwise. For the proposed framework, successful rejection required retrieval to stop when no sufficiently supported disease–aspect route could be established.

# 4 Results

The proposed framework was evaluated on the 577-question benchmark described in Section 3.11. Results are reported for disease routing, clinical-aspect routing, answer-retrieval fidelity, out-of-KG query handling, and explainability.

The proposed system was compared with five reimplemented published baselines described in Section 3.11: KG-based medical QAS [15], medical KBQA [16], QASTRisk [21], the cMedQA-based medical QA approach [2], and a recent hybrid knowledge-graph and large-language-model question-answering approach, denoted here as Hybrid KG–LLM QA [17]. GPT-4o, GPT-5.4, and MedGemma-27B-IT were also evaluated as zero-shot language-model baselines under the same benchmark and evaluation protocol.

## 4.1 Disease Routing Performance

Table 3 presents disease-routing performance across the four in-KG question categories. The proposed system achieved 100% across all metrics on straightforward questions and achieved 97% Recall@1, 98% MRR, and 99% nDCG@5 on long questions.

For spelling-error questions, GPT-4o and GPT-5.4 achieved the highest Recall@1 (98%), while the proposed system achieved 95% Recall@1, 100% Recall@5, 97% MRR, and 98% nDCG@5.

On MedQuAD, the proposed system achieved 96.61% Recall@1, 100% Recall@5, 97.46% MRR, and 98.07% nDCG@5. Hybrid KG–LLM QA achieved 94% Recall@1 on straightforward questions, 80% on long questions, 73% on spelling-error questions, and 81.92% on MedQuAD.

## 4.2 Clinical-Aspect Routing Performance

Clinical-aspect routing evaluates whether the system correctly identifies the requested type of clinical information. Table 4 reports performance across the four in-KG question categories.

The proposed system achieved 100% across all metrics for straightforward questions. For long questions, it achieved 95% Recall@1, 100% Recall@5, 97% MRR, and 98% nDCG@5. Hybrid KG–LLM QA performed notably lower on this category, with 21% Recall@1 and 57% Recall@5.

For spelling-error questions, the proposed system achieved 95% Recall@1, 100% Recall@5, 97% MRR, and 98% nDCG@5. Hybrid KG–LLM QA achieved 64% Recall@1 and 95% Recall@5.

On MedQuAD, the proposed system achieved 96.05% Recall@1, 100% Recall@5, 97.50% MRR, and 98.12% nDCG@5. Hybrid KG–LLM QA achieved 94.35% Recall@1 and 96.61% Recall@5, showing a much smaller gap than on long questions.

Overall, the largest difference between the two systems occurred on long questions, while their performance was closer on MedQuAD.

## 4.3 Answer Retrieval Fidelity

Answer retrieval fidelity was evaluated using BERTScore between each returned answer and the corresponding reference clinical content stored in the knowledge graph, following the evaluation procedure described in Section 3.11. In this evaluation, BERTScore measures semantic agreement with the expected KG evidence and is therefore interpreted as a measure of retrieval fidelity rather than as an independent assessment of clinical correctness.

Table 5 shows that the proposed system maintained consistently high retrieval fidelity across the four in-KG question categories. For straightforward questions, BERTScore reached 100% for precision, recall, and F1. This reflects the design of the proposed framework: when the disease and clinical aspect are routed correctly, the corresponding stored KG passage is returned directly. Because the same KG content is used as the reference answer, exact retrieval naturally produces complete agreement

**Table 3** Disease-routing performance across question types.

| Question Type | Approach | N | Recall@1 | Recall@5 | MRR | nDCG@5 |
|---|---|---|---|---|---|---|
| **Straightforward** | KG-based QAS | 100 | **100%** | **100%** | **100%** | **100%** |
| | KBQA | 100 | **100%** | **100%** | **100%** | **100%** |
| | QASTRisk | 100 | **100%** | **100%** | **100%** | **100%** |
| | cMedQA | 100 | **100%** | **100%** | **100%** | **100%** |
| | Hybrid KG–LLM QA | 100 | 94% | **100%** | 96% | 97% |
| | GPT-4o | 100 | 91% | 91% | 91% | 91% |
| | GPT-5.4 | 100 | 94% | 94% | 94% | 94% |
| | MedGemma-27B-IT | 100 | 66% | 92% | 74% | 78% |
| | **Proposed system** | 100 | **100%** | **100%** | **100%** | **100%** |
| **Long questions** | KG-based QAS | 100 | 94% | **100%** | 96% | 97% |
| | KBQA | 100 | 94% | **100%** | 96% | 97% |
| | QASTRisk | 100 | 88% | 90% | 88% | 89% |
| | cMedQA | 100 | 88% | 91% | 89% | 90% |
| | Hybrid KG–LLM QA | 100 | 80% | 87% | 82% | 83% |
| | GPT-4o | 100 | 92% | 98% | 94% | 95% |
| | GPT-5.4 | 100 | 93% | **100%** | 95% | 97% |
| | MedGemma-27B-IT | 100 | 47% | 87% | 56% | 63% |
| | **Proposed system** | 100 | **97%** | **100%** | **98%** | **99%** |
| **Spelling-error questions** | KG-based QAS | 100 | 87% | 91% | 88% | 89% |
| | KBQA | 100 | 90% | 98% | 92% | 93% |
| | QASTRisk | 100 | 94% | **100%** | 95% | 96% |
| | cMedQA | 100 | 92% | 96% | 94% | 94% |
| | Hybrid KG–LLM QA | 100 | 73% | 92% | 80% | 83% |
| | GPT-4o | 100 | **98%** | 98% | **98%** | **98%** |
| | GPT-5.4 | 100 | **98%** | 98% | **98%** | **98%** |
| | MedGemma-27B-IT | 100 | 95% | **100%** | 96% | 97% |
| | **Proposed system** | 100 | 95% | **100%** | 97% | **98%** |
| **MedQuAD user questions** | KG-based QAS | 177 | 88.70% | 90.40% | 89.31% | 89.58% |
| | KBQA | 177 | 95.48% | **100%** | 96.89% | 97.65% |
| | QASTRisk | 177 | 95.48% | 97.74% | 96.12% | 96.45% |
| | cMedQA | 177 | 93.79% | 98.87% | 95.54% | 96.31% |
| | Hybrid KG–LLM QA | 177 | 81.92% | 93.22% | 86.41% | 88.12% |
| | GPT-4o | 177 | 94.35% | 96.05% | 94.97% | 95.20% |
| | GPT-5.4 | 177 | 95.48% | 97.18% | 95.97% | 96.20% |
| | MedGemma-27B-IT | 177 | 79.66% | 92.66% | 83.83% | 85.99% |
| | **Proposed system** | 177 | **96.61%** | **100%** | **97.46%** | **98.07%** |

with the reference content. The 100% result therefore reflects exact retrieval fidelity to the stored KG evidence rather than independent clinical validation.

For long questions, BERTScore F1 decreased to 96.70%, while spelling-error questions achieved 96.20% and MedQuAD questions 97.91%. These reductions are consistent with the more difficult routing conditions observed for these categories. Because the retrieved clinical content is determined by the selected disease–aspect route, an incorrect disease or aspect can result in evidence that differs from the annotated KG reference even when the retrieval process itself remains controlled.

**Table 4** Clinical-aspect routing performance across question types.

| Question Type | Approach | N | Recall@1 | Recall@5 | MRR | nDCG@5 |
|---|---|---|---|---|---|---|
| **Straightforward** | KG-based QAS | 100 | **100%** | **100%** | **100%** | **100%** |
| | KBQA | 100 | 68% | 96% | 80% | 84% |
| | QASTRisk | 100 | 88% | 97% | 90% | 91% |
| | cMedQA | 100 | 88% | **100%** | 93% | 95% |
| | Hybrid KG–LLM QA | 100 | 85% | **100%** | 92% | 94% |
| | GPT-4o | 100 | **100%** | **100%** | **100%** | **100%** |
| | GPT-5.4 | 100 | **100%** | **100%** | **100%** | **100%** |
| | MedGemma-27B-IT | 100 | 94% | 98% | 95% | 96% |
| | **Proposed system** | 100 | **100%** | **100%** | **100%** | **100%** |
| **Long questions** | KG-based QAS | 100 | 78% | **100%** | 88% | 91% |
| | KBQA | 100 | 50% | 88% | 67% | 72% |
| | QASTRisk | 100 | 72% | 97% | 79% | 83% |
| | cMedQA | 100 | 74% | 99% | 84% | 87% |
| | Hybrid KG–LLM QA | 100 | 21% | 57% | 37% | 42% |
| | GPT-4o | 100 | 78% | 98% | 83% | 87% |
| | GPT-5.4 | 100 | 78% | **100%** | 83% | 87% |
| | MedGemma-27B-IT | 100 | 78% | 98% | 84% | 87% |
| | **Proposed system** | 100 | **95%** | **100%** | **97%** | **98%** |
| **Spelling-error questions** | KG-based QAS | 100 | 68% | 97% | 80% | 84% |
| | KBQA | 100 | 73% | 97% | 83% | 86% |
| | QASTRisk | 100 | 88% | **100%** | 91% | 93% |
| | cMedQA | 100 | 75% | 99% | 85% | 88% |
| | Hybrid KG–LLM QA | 100 | 64% | 95% | 75% | 80% |
| | GPT-4o | 100 | 91% | 98% | 94% | 95% |
| | GPT-5.4 | 100 | **98%** | 98% | **98%** | **98%** |
| | MedGemma-27B-IT | 100 | 95% | **100%** | 96% | 97% |
| | **Proposed system** | 100 | 95% | **100%** | 97% | **98%** |
| **MedQuAD user questions** | KG-based QAS | 177 | 93.79% | 96.05% | 94.74% | 94.89% |
| | KBQA | 177 | 95.48% | 96.05% | 96.67% | 95.76% |
| | QASTRisk | 177 | 89.27% | 98.87% | 92.31% | 93.90% |
| | cMedQA | 177 | 63.84% | 96.61% | 74.35% | 79.80% |
| | Hybrid KG–LLM QA | 177 | 94.35% | 96.61% | 95.15% | 95.51% |
| | GPT-4o | 177 | 88.14% | 97.74% | 91.40% | 92.97% |
| | GPT-5.4 | 177 | 91.53% | 94.92% | 92.85% | 93.06% |
| | MedGemma-27B-IT | 177 | 82.49% | 95.48% | 87.26% | 89.32% |
| | **Proposed system** | 177 | **96.05%** | **100%** | **97.50%** | **98.12%** |

A different pattern is observed for Hybrid KG–LLM QA. Its BERTScore F1 was 39.98% for straightforward questions, 17.55% for long questions, 27.33% for spelling-error questions, and 47.44% for MedQuAD. The particularly low value for long questions is consistent with its weaker clinical-aspect routing on that category. In addition, because its final answer is generated from retrieved evidence, differences in wording, compression, or content selection can further reduce similarity with the stored reference answer.

These results therefore reflect both routing quality and the way in which retrieved evidence is presented. For the proposed system, the accepted disease–aspect route directly determines the stored clinical content returned to the user, whereas generative

**Table 5** Answer retrieval fidelity measured using BERTScore across the four in-KG question categories.

| Question Type | Approach | N | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **Straightforward** | KG-based QAS | 100 | 98.59% | 98.59% | 98.58% |
| | KBQA | 100 | 74.27% | 78.59% | 76.37% |
| | QASTRisk | 100 | 91.59% | 91.59% | 91.58% |
| | cMedQA | 100 | 91.17% | 96.84% | 93.79% |
| | Hybrid KG–LLM QA | 100 | 39.97% | 39.99% | 39.98% |
| | GPT-4o | 100 | 85.80% | 82.18% | 83.82% |
| | GPT-5.4 | 100 | 89.33% | 92.82% | 90.90% |
| | MedGemma-27B-IT | 100 | 69.18% | 69.28% | 69.19% |
| | **Proposed system** | 100 | **100.00%** | **100.00%** | **100.00%** |
| **Long questions** | KG-based QAS | 100 | 87.38% | 87.42% | 87.39% |
| | KBQA | 100 | 60.48% | 62.52% | 61.48% |
| | QASTRisk | 100 | 82.48% | 82.52% | 82.49% |
| | cMedQA | 100 | 91.24% | 96.01% | 93.46% |
| | Hybrid KG–LLM QA | 100 | 19.77% | 15.78% | 17.55% |
| | GPT-4o | 100 | 85.34% | 82.49% | 83.76% |
| | GPT-5.4 | 100 | 79.33% | 82.11% | 80.54% |
| | MedGemma-27B-IT | 100 | 60.92% | 60.42% | 60.66% |
| | **Proposed system** | 100 | **96.71%** | **96.70%** | **96.70%** |
| **Spelling-error questions** | KG-based QAS | 100 | 83.11% | 83.06% | 83.06% |
| | KBQA | 100 | 83.11% | 83.06% | 83.06% |
| | QASTRisk | 100 | 83.11% | 83.06% | 83.06% |
| | cMedQA | 100 | 91.28% | 95.94% | 93.48% |
| | Hybrid KG–LLM QA | 100 | 31.81% | 23.95% | 27.33% |
| | GPT-4o | 100 | 87.65% | 82.45% | 84.91% |
| | GPT-5.4 | 100 | 81.30% | 84.06% | 82.61% |
| | MedGemma-27B-IT | 100 | 95.93% | 95.88% | 95.89% |
| | **Proposed system** | 100 | **96.40%** | **96.00%** | **96.20%** |
| **MedQuAD user questions** | KG-based QAS | 177 | 91.25% | 91.33% | 91.29% |
| | KBQA | 177 | 92.46% | 92.54% | 92.50% |
| | QASTRisk | 177 | 91.35% | 91.43% | 91.39% |
| | cMedQA | 177 | 87.68% | 92.63% | 89.96% |
| | Hybrid KG–LLM QA | 177 | 48.07% | 46.82% | 47.44% |
| | GPT-4o | 177 | 86.44% | 81.87% | 84.03% |
| | GPT-5.4 | 177 | 89.89% | 92.53% | 91.15% |
| | MedGemma-27B-IT | 177 | 74.29% | 74.33% | 74.30% |
| | **Proposed system** | 177 | **97.70%** | **98.13%** | **97.91%** |

approaches may transform the retrieved evidence before presenting the final answer. BERTScore should consequently be interpreted specifically as a measure of answer-retrieval fidelity and semantic agreement with the KG reference content, rather than as an independent measure of clinical correctness.

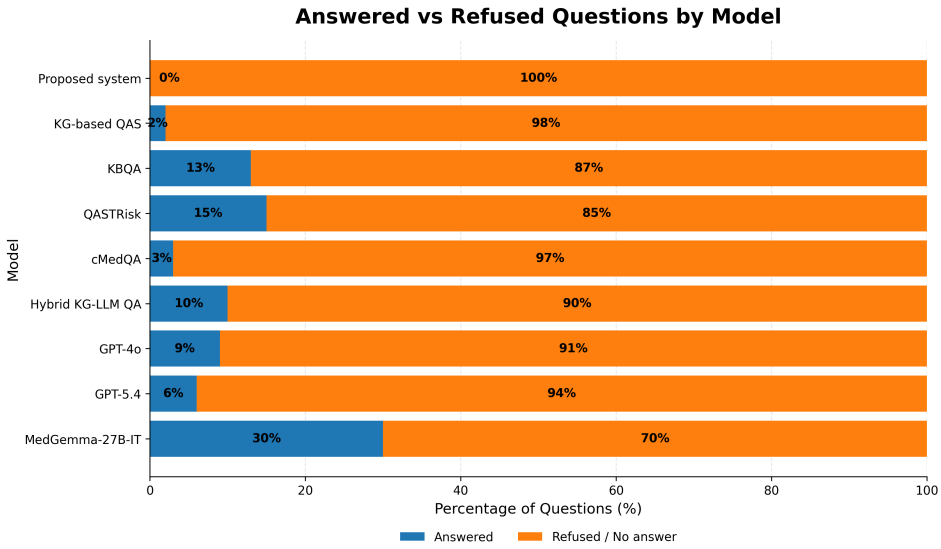


**Fig. 6** Out-of-KG query-handling results for 100 questions whose target conditions were absent from the cardiovascular knowledge graph. "Answered" indicates that medical content was returned for an unsupported condition, whereas "Refused / No answer" indicates that no unsupported medical answer was returned.

## 4.4 Out-of-KG Query Handling

Out-of-KG evaluation assessed whether each approach avoided returning medical content for conditions that were absent from the cardiovascular knowledge graph. This constitutes the operational form of out-of-distribution evaluation used in this study. A total of 100 out-of-KG questions were evaluated. The questions included symptoms and clinical terminology that could overlap with supported cardiovascular conditions, creating the possibility of incorrectly mapping an unsupported condition to a related in-KG disease.

For this evaluation, *answered* indicates that medical content was returned for an unsupported condition, whereas *refused/no answer* indicates that no unsupported medical answer was returned. The latter category does not imply that all approaches implemented the same explicit refusal mechanism. For the zero-shot language-model baselines, the evaluation prompt explicitly instructed the models not to map unsupported conditions to the closest in-KG disease and not to generate medical content for out-of-KG questions.

As shown in Figure 6, the proposed system returned no unsupported medical answer for all 100 out-of-KG questions. KG-based QAS did so for 98 questions, cMedQA for 97, GPT-5.4 for 94, GPT-4o for 91, Hybrid KG–LLM QA for 90, KBQA for 87, QASTRisk for 85, and MedGemma-27B-IT for 70. Correspondingly, Hybrid KG–LLM QA returned medical content for 10 unsupported questions, while MedGemma-27B-IT showed the highest number of answered out-of-KG questions at 30.

In the proposed system, a route that does not satisfy the predefined 92.5% routing threshold is subjected to T5-based recovery, in which three reformulations are generated and independently routed, as described in Section 3.8. If none of the reformulations satisfies the routing criterion, knowledge-graph retrieval is stopped and the system returns:

> "Warning: We don't have a satisfactory answer for this question despite several paraphrasing attempts."

The proposed system therefore did not force any of the 100 evaluated out-of-KG questions onto an accepted disease–aspect route. This result is specific to the out-of-KG benchmark evaluated in this study and should not be interpreted as evidence of universal out-of-distribution detection.

## 4.5 Error Breakdown Analysis

Figure 7 jointly evaluates disease and clinical-aspect routing using four outcomes: both correct, disease correct but aspect incorrect, aspect correct but disease incorrect, and both incorrect. For multi-request questions, "both correct" required all annotated disease–aspect routes to be recovered correctly.

The proposed system achieved joint correctness of 100% on straightforward questions, 95% on long questions, 95% on spelling-error questions, and 96.05% on MedQuAD. For long questions, the remaining errors mainly involved incorrect aspect routing or errors in both components.

The comparison also shows that some baselines preserved the correct disease but selected the wrong clinical aspect, particularly for long questions. This pattern was most pronounced for Hybrid KG–LLM QA. Overall, the analysis complements the separate Recall@1 results by showing whether routing errors arise from the disease, the aspect, or both.

## 4.6 Explainability Performance: CTRA versus LIME

The proposed Causal Token Role Attribution (CTRA) method was compared with LIME [22] on the same deployed disease–aspect router. As described in Section 3.10, LIME was connected to the router through a two-class route-preservation wrapper, so each perturbed query was evaluated according to whether it preserved or changed the accepted disease–aspect route. LIME explained the *route preserved* class, while CTRA measured changes directly from the router's continuous disease- and aspect-support profiles. Both methods therefore explained the same accepted routing decisions, although they used different attribution mechanisms.

LIME was configured with 300 perturbation samples per query and a fixed random seed of 42. The standard `LimeTextExplainer` settings were retained for parameters not explicitly specified. Up to 12 query features were exposed to the local surrogate model. Positively weighted features for the *route preserved* class were ranked by the magnitude of their LIME weights, and up to the five highest-ranked positive words were retained as the explanation. If fewer than five positively weighted words were available, all available positive words were retained.

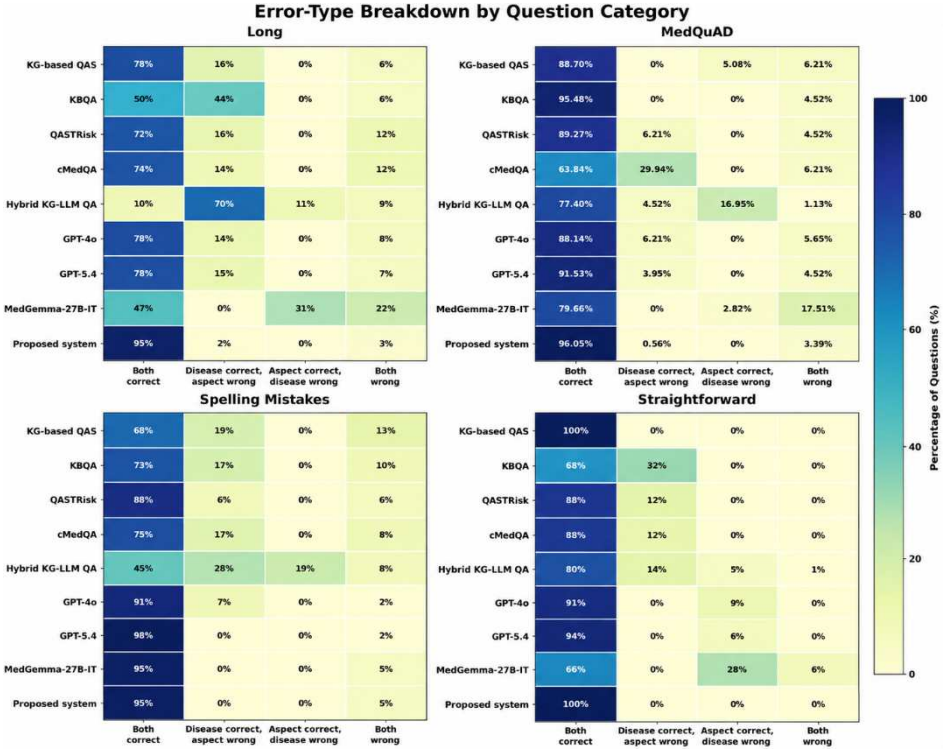


**Fig. 7** Joint disease–aspect routing error breakdown across the four in-KG question categories.

Both methods were evaluated using the same sufficiency, consistency, explanation-length, and runtime definitions described in Section 3.10. Explanation consistency was calculated only for questions with at least two distinct route-preserving T5 alternatives, as defined in that section. The value of $N$ in Table 6 denotes the total number of questions evaluated for sufficiency, explanation length, and runtime, whereas $N_{\text{cons}}$ denotes the number contributing to the consistency calculation. Runtime measured explanation generation only and excluded T5 alternative generation, rerouting of alternative formulations, and subsequent calculation of the evaluation measures.

As shown in Table 6, CTRA produced higher sufficiency and consistency values than LIME across the four evaluated question categories. The difference in sufficiency was most pronounced for long questions, increasing from 68.00% with LIME to 90.00% with CTRA. For straightforward questions, consistency increased from 74.08% to 96.39%, while on MedQuAD it increased from 53.89% to 67.00%.

CTRA also produced explanations that were equal in length or shorter than those produced by LIME. For straightforward questions, both methods selected an average of 2.97 words. For long, spelling-error, and MedQuAD questions, CTRA used fewer explanation words while retaining higher sufficiency.

A substantial difference was also observed in runtime. LIME required approximately 42–44 seconds per query, whereas CTRA required 3.89–7.46 seconds per query

**Table 6** Comparison of LIME and CTRA for explanation quality and efficiency. $N$ is the total number of evaluated questions, and $N_{\text{cons}}$ is the number of questions included in the consistency calculation.

| Category | Method | N | $N_{\text{cons}}$ | Sufficiency (%) | Consistency (%) | Avg. Length (words) | Avg. Runtime (s/query) |
|---|---|---|---|---|---|---|---|
| **Straightforward** | LIME | 100 | 100 | 66.00 | 74.08 | **2.97** | 42.39 |
| | **Proposed CTRA** | 100 | 100 | **83.00** | **96.39** | **2.97** | **4.44** |
| **Long** | LIME | 100 | 100 | 68.00 | 96.49 | 3.79 | 42.48 |
| | **Proposed CTRA** | 100 | 100 | **90.00** | **98.38** | **2.99** | **7.46** |
| **Spelling-error questions** | LIME | 100 | 100 | 58.00 | 81.94 | 2.98 | 43.91 |
| | **Proposed CTRA** | 100 | 100 | **61.00** | **84.81** | **2.64** | **4.59** |
| **MedQuAD user questions** | LIME | 177 | 177 | 77.97 | 53.89 | 2.81 | 43.17 |
| | **Proposed CTRA** | 177 | 177 | **80.97** | **67.00** | **2.62** | **3.89** |

across the four categories. This difference is consistent with LIME evaluating 300 perturbed samples and fitting a local surrogate model for each query, whereas CTRA measures the effect of controlled word-removal interventions directly on the deployed router.

Overall, the results show different behaviour from the two explanation mechanisms. LIME estimates word importance from perturbed samples using a locally fitted surrogate model for preservation of the accepted route, whereas CTRA directly measures how controlled word removals change the disease- and aspect-support profiles of the deployed router. CTRA produced more sufficient and consistent explanations in this evaluation while requiring substantially less runtime.

## 4.7 Ablation Study

An ablation study was conducted to assess the contribution of spelling correction, controlled query segmentation, and T5-based low-support recovery. Each component was removed independently while all other system and evaluation settings were kept unchanged. Disease and clinical-aspect routing were evaluated using Recall@1, Recall@5, MRR, and nDCG@5.

The analysis included 100 straightforward questions, 100 long questions, 100 spelling-error questions, and 177 selected MedQuAD questions. The 100 out-of-KG questions were excluded because their target conditions are intentionally absent from the knowledge graph and are evaluated separately through refusal performance in Section 4.4.

The ablation results show that the three components address different forms of query difficulty.

Removing spelling correction affected only the spelling-error subset. Disease Recall@1 decreased from 95% to 89%, while clinical-aspect Recall@1 decreased from 95% to 62%. Although the spelling errors were introduced in disease expressions, aspect routing is also affected because the semantic aspect router represents the complete processed query. An uncorrected disease expression can therefore alter the query representation used to distinguish between aspect candidates, particularly when an explicit lexical aspect cue is insufficient. This explains why spelling correction improves both disease and aspect routing on this subset.

Controlled segmentation had its strongest effect on long questions. Without segmentation, disease Recall@1 decreased from 97% to 90%, while clinical-aspect

**Table 7** Ablation study evaluating the contribution of spelling correction, query segmentation, and T5 recovery to disease and clinical-aspect routing performance. R@1 and R@5 denote Recall@1 and Recall@5, respectively.

| | | Disease Routing | | | | Clinical-Aspect Routing | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Question Type** | **Configuration** | **R@1** | **R@5** | **MRR** | **nDCG@5** | **R@1** | **R@5** | **MRR** | **nDCG@5** |
| **Straightforward** | **Full system** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| | No spelling correction | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | No segmentation | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | No T5 recovery | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **Long** | **Full system** | **97** | **100** | **98** | **99** | **95** | **100** | **97** | **98** |
| | No spelling correction | 97 | 100 | 98 | 99 | 95 | 100 | 97 | 98 |
| | No segmentation | 90 | 98 | 92 | 94 | 83 | 100 | 89 | 92 |
| | No T5 recovery | 94 | 98 | 95 | 96 | 93 | 100 | 95 | 97 |
| **Spelling-error** | **Full system** | **95** | **100** | **97** | **98** | **95** | **100** | **97** | **98** |
| | No spelling correction | 89 | 89 | 89 | 89 | 62 | 80 | 66 | 70 |
| | No segmentation | 95 | 100 | 97 | 98 | 95 | 100 | 97 | 98 |
| | No T5 recovery | 95 | 100 | 97 | 98 | 95 | 100 | 97 | 98 |
| **MedQuAD** | **Full system** | **96.61** | **100** | **97.46** | **98.07** | **96.05** | **100** | **97.50** | **98.12** |
| | No spelling correction | 96.61 | 100 | 97.46 | 98.07 | 96.05 | 100 | 97.50 | 98.12 |
| | No segmentation | 93.79 | 100 | 95.50 | 96.59 | 94.92 | 100 | 96.30 | 97.19 |
| | No T5 recovery | 96.61 | 100 | 97.46 | 98.07 | 96.05 | 100 | 97.50 | 98.12 |

Recall@1 decreased from 95% to 83%. MRR and nDCG@5 also decreased for both tasks. Clinical-aspect Recall@5 remained at 100%, showing that the correct aspect was generally still retrieved but was less consistently ranked first. This supports the role of segmentation in separating multiple or structurally complex requests before routing.

A smaller segmentation effect was observed on MedQuAD. Disease Recall@1 decreased from 96.61% to 93.79%, and clinical-aspect Recall@1 decreased from 96.05% to 94.92%, while Recall@5 remained at 100% for both tasks. Segmentation therefore also improved top-ranked routing for a subset of the external questions.

Removing T5 recovery affected only the long-question subset. Disease Recall@1 decreased from 97% to 94%, and clinical-aspect Recall@1 decreased from 95% to 93%. This is consistent with T5 being used only when the initial disease–aspect route falls below the predefined support threshold; in this benchmark, its measurable benefit occurred on low-support long questions.

Straightforward questions remained unchanged under all ablations, with 100% performance across all routing metrics. These questions were therefore sufficiently direct for the base routing procedure without requiring additional correction, segmentation, or recovery to improve the measured results.

Overall, the ablation study shows complementary contributions: spelling correction improves handling of lexical errors, controlled segmentation improves routing of complex and multi-request questions, and T5 recovery assists low-support routes. The full system combines these mechanisms to address different sources of routing difficulty.

# 5 Discussion

The results show that the strength of the proposed framework lies in the integration of explicit disease–aspect routing, controlled knowledge-graph retrieval, unsupported-query handling, source-linked evidence, and route-level explanation. Language models and pretrained semantic encoders support query interpretation and reformulation, whereas the final medical content is retrieved directly from the accepted knowledge-graph route rather than freely generated.

The routing results support the proposed design, particularly for complex questions. For long questions, the system achieved 97% disease Recall@1 and 95% clinical-aspect Recall@1. On MedQuAD, Recall@1 reached 96.61% for disease routing and 96.05% for aspect routing, with the highest MRR and nDCG@5 values.

These results support separating disease identification from clinical-aspect identification. Long and multi-clause questions may contain several requests, which can be obscured when the complete question is treated as a single routing decision. The proposed framework addresses this through controlled segmentation and independent disease–aspect routing. It also combines complementary evidence, using exact, fuzzy, and semantic matching for diseases and lexical and semantic evidence for aspects, reducing dependence on any single routing signal. The joint error analysis further showed that, for long questions, some competing approaches identified the disease correctly while failing to preserve the requested clinical aspect, whereas this error type was substantially reduced in the proposed framework.

For spelling-error questions, GPT-4o and GPT-5.4 achieved higher disease Recall@1, while GPT-5.4 also achieved higher clinical-aspect Recall@1. Nevertheless, the proposed system retained 100% Recall@5 and 98% nDCG@5 for both routing tasks. This indicates that the correct disease and aspect generally remained among the highest-ranked candidates despite occasional errors at rank one. Routing performance should also be considered alongside the controlled retrieval process. Accepted routes retrieve source-linked evidence directly from the knowledge graph, while insufficiently supported routes are withheld when recovery fails.

The answer-retrieval results further reflect the importance of the selected disease–aspect route. The proposed system achieved the highest BERTScore F1 across all four in-KG question categories. Because the accepted route determines the clinical content retrieved from the knowledge graph, an incorrect disease or aspect results in different evidence being returned. Generative approaches may also paraphrase, compress, or omit retrieved information, whereas the proposed framework returns the stored KG content directly. BERTScore is therefore interpreted as fidelity to the KG reference content rather than as a measure of overall clinical answer quality.

This distinction was particularly visible for Hybrid KG–LLM QA. Although its disease-routing performance remained substantially higher than its aspect-routing performance on long questions, it also showed lower answer-retrieval fidelity. This pattern supports the importance of preserving individual requests in multi-clause questions and separating disease and aspect identification before evidence retrieval.

The out-of-KG evaluation examined whether systems returned medical information for conditions absent from the knowledge graph. The proposed framework returned no unsupported medical answer for any of the 100 evaluated out-of-KG questions. Queries

**Table 8** Methodological comparison of the proposed system with the published baseline approaches evaluated in this study. Out-of-KG query-handling values refer to behaviour observed on the 100-question out-of-KG benchmark used in this work.

| Dimension | KG-based QAS [15] | KBQA [16] | cMedQA [2] | QASTRisk [21] | Hybrid KG–LLM QA [17] | Proposed system |
|---|---|---|---|---|---|---|
| **Primary domain** | Medical (general) | Medical (general) | Medical (general) | Risk / engineering | Diabetes | Cardiovascular medicine |
| **Query matching / routing** | TF–IDF with Naïve Bayes intent classification | Entity and relation extraction with intent recognition and weighted path ranking | Elasticsearch candidate retrieval with pretrained-language-model semantic matching and KG/NER-enhanced ranking | GPT-based query interpretation and graph-query construction | Qwen2.5-based entity recognition, BERT–CNN intent classification, and intent-to-KG relation mapping | PubMedBERT, BioBERT, Sentence Transformer, lexical matching, and explicit disease–aspect routing |
| **Final answer** | Retrieved KG snippets | Candidate answers inferred from the medical KG | Retrieved candidate answer selected through semantic matching and ranking | LLM-generated answer from KG query output | LLM-generated answer from KG evidence and RAG-retrieved guideline passages | Retrieved content from the accepted disease–aspect KG route |
| **Generative model used for final answer** | No | No | No | Yes | Yes | No |
| **Explicit disease–aspect separation** | Not explicitly reported | Not explicitly reported | Not explicitly reported | Not explicitly reported | Not explicitly reported | Yes |
| **Out-of-KG query handling** | 98/100 no unsupported answer | 87/100 no unsupported answer | 97/100 no unsupported answer | 85/100 no unsupported answer | 90/100 no unsupported answer | **100/100 explicit safe refusal** |
| **Source-quality assessment** | Not reported | Not reported | Not reported | Not reported | Not reported | Adapted CRAAP-based source-quality assessment retained with the evidence |
| **Evidence provenance** | KG-based evidence | KG-based evidence | KG-based evidence used in answer retrieval and ranking | KG query output | KG evidence and retrieved guideline passages exposed for traceability | Source and reference information retained in the KG |
| **Routing explanation** | Not explicitly integrated | Not explicitly integrated | Not explicitly integrated | Not explicitly integrated | Intermediate entity, intent, and evidence outputs exposed; no word-level routing attribution | CTRA explains how individual words affect the deployed disease–aspect router |

below the 92.5% routing threshold underwent T5-based reformulation, and retrieval was stopped when none of the reformulations established a sufficiently supported route. This demonstrates effective rejection within the evaluated out-of-KG set, but it should not be interpreted as universal out-of-distribution detection.

Table 8 compares the principal methodological characteristics of the proposed framework with the five published baseline approaches evaluated in this study.

The comparison highlights two main methodological distinctions. First, the proposed framework explicitly separates disease identification from clinical-aspect identification before retrieval, whereas the evaluated baselines do not report the same disease–aspect routing structure. Second, generation is deliberately restricted. T5 is used only to reformulate low-support queries, while an accepted disease–aspect route directly determines the knowledge-graph content returned to the user. The design objective is therefore tighter control over what medical evidence can be retrieved and where that evidence originates.

Recent research has explored different roles for language models in conjunction with structured knowledge, including faithful reasoning over knowledge graphs [19] and language-model-assisted knowledge-graph construction [20]. Within the proposed framework, language models support query interpretation and low-support reformulation, while the final medical evidence is retrieved directly from the accepted disease–aspect route in the knowledge graph. Because prior work addresses different

tasks and evaluation settings, the comparison is methodological rather than a direct comparison of numerical performance.

The source-quality component should also be distinguished from model confidence and retrieval ranking. The adapted CRAAP assessment evaluates characteristics of the source associated with the retrieved evidence and does not estimate the probability that an answer is correct. Retaining this assessment together with source provenance and references improves traceability without presenting source quality as prediction confidence.

The explainability results provide a further distinction. CTRA achieved higher explanation sufficiency and consistency than LIME across all evaluated categories while requiring substantially less runtime. Unlike LIME, which estimates feature importance using a locally fitted surrogate model, CTRA removes words and reruns the deployed disease–aspect router to measure their effect on the routing decision. The explanation is therefore directly aligned with the mechanism that determines subsequent knowledge-graph retrieval. The term *causal* refers only to this controlled input intervention and does not imply biological or clinical causality.

The findings should nevertheless be interpreted within the scope of the evaluation. The benchmark contains 577 questions and is restricted to cardiovascular conditions represented in the constructed knowledge graph. Four benchmark categories were manually constructed around the supported disease–aspect schema. Although these questions were medically reviewed and an external MedQuAD subset was included, evaluation on larger independently constructed benchmarks is required to assess broader generalisability. The 100-question out-of-KG set does not represent every possible unsupported medical condition, and BERTScore measures semantic agreement with the reference content rather than independent clinical correctness. CTRA also showed lower explanation consistency on MedQuAD than on the other evaluated question categories.

Overall, the framework changes the role of language models in medical question answering. Rather than generating the final medical response, language models are restricted to query interpretation and low-support reformulation, while the accepted disease–aspect route determines which curated evidence is retrieved. Weakly supported routes can be rejected, whereas accepted routes return source-linked knowledge-graph content and expose the reasoning behind the routing decision. This preserves the origin of the returned evidence and prevents the final medical content from being produced through unconstrained generation.

# 6 Conclusion

This work presented a reliable cardiovascular question-answering framework in which language models are used to interpret and route a question rather than to generate the final medical response. Each query is mapped to a disease–aspect route using adaptive lexical and semantic evidence (Section 3.7), and the strength of that route is assessed before any clinical content is returned (Section 3.8). Accepted routes retrieve the corresponding source-linked evidence directly from the cardiovascular knowledge graph (Sections 3.3 and 3.9), whereas insufficiently supported routes are withheld

rather than forced onto the nearest available condition. CTRA then exposes which words in the accepted query formulation influenced the routing decision (Section 3.10). The resulting design therefore separates language understanding from medical-content generation, so that the evidence returned to the user remains traceable to a defined knowledge-graph route and its underlying source.

The evaluation shows that the proposed framework maintained high routing accuracy across different forms of query difficulty. For long questions, disease Recall@1 reached 97% and clinical-aspect Recall@1 reached 95% (Sections 4.1 and 4.2). This routing performance translated into consistently high answer-retrieval fidelity, with the proposed system achieving the highest BERTScore F1 across all four in-KG question categories (Section 4.3). The framework returned no unsupported medical answer for any of the 100 out-of-KG questions (Section 4.4). Finally, CTRA produced more sufficient and more consistent explanations than LIME across all evaluated categories and reduced explanation runtime substantially (Section 4.6).

The main contribution of this work is a controlled QA pipeline in which disease–aspect routing (Section 3.7), route validation and rejection (Section 3.8), source-linked knowledge-graph retrieval (Sections 3.3 and 3.9), and routing explanation (Section 3.10) remain explicitly connected. As a result, the final medical content is retrieved from an accepted and traceable knowledge-graph route rather than freely generated.

The present study is limited to a cardiovascular knowledge graph and a 577-question benchmark. Cardiovascular medicine was selected as a focused test domain because it provides a clinically coherent setting with diverse disease–aspect information needs and access to established public sources such as the NHS, BHF, and MedlinePlus. Importantly, the main methodological components are not specific to cardiovascular medicine. The proposed system can be applied to other domains by changing the supported entities, aspects, and knowledge-graph schema.

Future work will therefore extend the framework to additional medical specialties and larger independently constructed benchmarks, with broader and more adversarial out-of-KG evaluation. Further directions include multilingual question answering, automatic updating of source-linked knowledge as clinical guidance changes, evaluation under conflicting or incomplete source evidence, improved calibration of route acceptance, and prospective clinician assessment of both retrieved evidence and routing explanations.

The central finding is that reliable medical question answering does not require the final response to be freely generated. By admitting only sufficiently supported routes and retrieving the corresponding source-linked evidence directly from the knowledge graph, the framework maintains a verifiable connection between the user's question and the medical information returned, while CTRA makes the routing decision itself inspectable (Sections 3.8, 3.9, and 3.10).

## Supplementary information

Supplementary information is available with this article and includes the following materials:

- **Supplementary File 1**: Source Quality Assessment Framework.
- **Supplementary File 2**: Example System Outputs and XAI Visualisations.


## Acknowledgements

The authors would like to thank the Medical College Research Centre at King Khalid University for supporting the medical expert review process. We also thank the three participating medical experts for their contributions to the source-quality assessment and review of the manually constructed benchmark and its annotations.


## Declarations


**Funding**
This research was supported by a PhD scholarship from the Cultural Bureau, Embassy of Saudi Arabia (SACB). The funder had no role in the study design, data collection, analysis, interpretation, or decision to submit the manuscript for publication.


**Conflict of interest**
The authors declare no competing interests.

**Ethics declarations**
This study did not involve human participants, biological samples, private data, or personally identifiable information. The system was developed using publicly accessible health-education content from the National Health Service (NHS), British Heart Foundation (BHF), and MedlinePlus [9–11]. Therefore, informed consent was not required, and no individual-level or sensitive data were collected or stored.

Disease-related images were included only where appropriate provenance and usage information were available. Image source, creator or contributor, and licence or usage information were retained, and no identifiable human images were included.

The system is intended for educational and informational purposes only and does not provide diagnosis, treatment decisions, or personalised medical advice. Responses are restricted to curated knowledge-graph content, include source and source-quality information, and are withheld when a supported disease–aspect route cannot be established. Users should consult qualified healthcare professionals for individual medical concerns.

**Consent for publication**
Not applicable.

**Clinical Trial Number**
Not applicable.

**Consent to Participate declaration**
Not applicable.

**Data availability**
The clinical content used in this study is publicly available from the NHS, BHF, and MedlinePlus websites [9–11]. Medical images were obtained from Radiopaedia and Wikimedia Commons [28]. The evaluation benchmark and related materials are available through the GitHub repository.

**Code availability**
The source code, evaluation benchmark, selected MedQuAD questions, zero-shot prompts, and model settings used in this study are available in the accompanying GitHub repository: https://github.com/mualshamri/Explainable-Cardiovascular-Question-Answering.git

**Author contributions**
Mohammed Alshammari and Amir Atapour-Abarghouei equally contributed to conceptualisation. Mohammed Alshammari and Amir Atapour-Abarghouei collaboratively performed the methodology, benchmark annotation, and validation. Software development, data curation, formal analysis, and the writing of the original draft were handled by Mohammed Alshammari. Finally, the writing—review and editing were carried out by both Mohammed Alshammari and Amir Atapour-Abarghouei, with supervision provided by Amir Atapour-Abarghouei.

# References


[1] Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal **8**(2), 188–194 (2021) https://doi.org/10.7861/fhj.2021-0095

[2] Guo, Q., Cao, S., Yi, Z.: A medical question answering system using large language models and knowledge graphs. International Journal of Intelligent Systems **37**(11), 8548–8564 (2022) https://doi.org/10.1002/int.22955

[3] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023) https://doi.org/10.1145/3571730

[4] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2) (2025) https://doi.org/10.1145/3703155

[5] Banerjee, S., Alsop, P., Jones, L., Cardinal, R.N.: Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. Patterns **3**(6), 100506 (2022) https://doi.org/10.1016/j.patter.2022.100506

[6] Murali, L., Gopakumar, G., Viswanathan, D.M., Nedungadi, P.: Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. Journal of Biomedical Informatics **143**, 104403 (2023) https://doi.org/10.1016/j.jbi.2023.104403

[7] Abu-Salih, B., AL-Qurishi, M., Alweshah, M., *et al.*: Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. Journal of Big Data **10**, 81 (2023) https://doi.org/10.1186/s40537-023-00774-9

[8] Wood, E.C., Glen, A.K., Kvarfordt, L.G., *et al.*: RTX-KG2: A system for building a semantically standardized knowledge graph for translational biomedicine. BMC Bioinformatics **23**, 400 (2022) https://doi.org/10.1186/s12859-022-04932-3

[9] National Health Service: Health A to Z. https://www.nhs.uk/health-a-to-z/. Accessed August 22, 2026 (2026)

[10] British Heart Foundation: Heart and Circulatory Conditions. https://www.bhf.org.uk/informationsupport/conditions. Accessed August 22, 2026 (2026)

[11] MedlinePlus: Health Topics. https://medlineplus.gov/healthtopics.html. Accessed August 22, 2026 (2026)

[12] Blakeslee, S.: The craap test. LOEX Quarterly **31**(3) (2004). Article 4

[13] Sye, D., Thompson, D.: Tools, tests, and checklists: The evolution and future of source evaluation frameworks. Journal of New Librarianship **8**(1), 76–100 (2023) https://doi.org/10.33011/newlibs/13/9

[14] Simms, A.M., Kanakia, A., Sipra, M., *et al.*: A patient safety knowledge graph supporting vaccine product development. BMC Medical Informatics and Decision Making **24**, 10 (2024) https://doi.org/10.1186/s12911-023-02409-8

[15] Zhang, K.: A knowledge graph based medical intelligent question answering system. In: 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, pp. 460–465 (2021). https://doi.org/10.1109/CEI52496.2021.9574515

[16] Huang, X., Zhang, J., Xu, Z., Ou, L., Tong, J.: A knowledge graph based question answering method for medical domain. PeerJ Computer Science **7**, 667 (2021) https://doi.org/10.7717/peerj-cs.667

[17] Zhang, Y., Li, S., Lyu, T., Yu, X., Jiang, T., Li, X., He, G., Zhou, W.: Diabetes question answering method based on knowledge graph and large language model. Discover Artificial Intelligence **6**, 1206 (2026) https://doi.org/10.1007/s44163-026-02078-2

[18] Gao, J., Cao, J., Bu, R., Zhu, N., Guan, W., Yu, H.: Promoting knowledge base question answering by directing llms to generate task-relevant logical forms. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 23914–23922 (2025). https://doi.org/10.1609/aaai.v39i22.34564 . https://doi.org/10.1609/aaai.v39i22.34564

[19] Sui, Y., He, Y., Liu, N., He, X., Wang, K., Hooi, B.: FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 8315–8330. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653

/v1/2025.findings-acl.436 . https://aclanthology.org/2025.findings-acl.436/

[20] Li, J., Yu, H., Ma, Z., Qi, P.: Discovering latent facts from context to construct richer open knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 19233–19241 (2026). https://doi.org/10.1609/aaai.v40i23.38998 . https://doi.org/10.1609/aaai.v40i23.38998

[21] Isah, M.A., Kim, B.-S.: Question-answering system powered by knowledge graph and generative pretrained transformer to support risk identification in tunnel projects. Journal of Construction Engineering and Management **151**(1), 04024193 (2025) https://doi.org/10.1061/JCEMD4.COENG-15230

[22] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778 . https://doi.org/10.1145/2939672.2939778

[23] DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized nlp models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.408 . https://aclanthology.org/2020.acl-main.408/

[24] Huang, G., Li, Y., Jameel, S., Long, Y., Papanastasiou, G.: From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? Computational and Structural Biotechnology Journal **24**, 362–373 (2024) https://doi.org/10.1016/j.csbj.2024.05.004

[25] Hajba, G.L.: Website Scraping with Python: Using BeautifulSoup and Scrapy, 1st edn. Apress, Berkeley, CA (2018). https://doi.org/10.1007/978-1-4842-3925-4 . https://doi.org/10.1007/978-1-4842-3925-4

[26] Chandraprabha, Kumar, A., Saxena, S.: Data driven testing framework using selenium webdriver. International Journal of Computer Applications **118**(18), 18–23 (2015) https://doi.org/10.5120/20845-3497

[27] RadPPTX: Image Downloader for Radiopaedia. Chrome Web Store extension, Version 1.2.0. Accessed 2026 (2026). https://chromewebstore.google.com/detail/image-downloader-for-radi/ijiffbpdcojpknpgcoallehenpikkdpd

[28] Wikimedia Commons: Wikimedia Commons. https://commons.wikimedia.org/. Accessed: 28 August 2026 (2026)

[29] Webber, J.: A programmatic introduction to neo4j. In: Proceedings of the 3rd

Annual Conference on Systems, Programming, and Applications: Software for Humanity. SPLASH '12, pp. 217–218. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2384716.2384777 . https://doi.org/10.1145/2384716.2384777

[30] Omelianchuk, K., Atrasevych, V., Chernodub, A., Skurzhanskyi, O.: GECToR – grammatical error correction: Tag, not rewrite. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 163–170. Association for Computational Linguistics, Seattle, WA, USA – Online (2020). https://doi.org/10.18653/v1/2020.bea-1.16 . https://aclanthology.org/2020.bea-1.16/

[31] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1) (2021) https://doi.org/10.1145/3458754

[32] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020) https://doi.org/10.1093/bioinformatics/btz682

[33] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410 . https://aclanthology.org/D19-1410/

[34] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)

[35] Liu, J.: LlamaIndex. Software. https://github.com/run-llama/llama_index

[36] Ben Abacha, A., Demner-Fushman, D.: A question-entailment approach to question answering. BMC Bioinformatics **20**(1), 511 (2019) https://doi.org/10.1186/s12859-019-3119-4

[37] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008). https://doi.org/10.1017/CBO9780511809071 . https://doi.org/10.1017/CBO9780511809071

[38] Voorhees, E.M., Tice, D.M.: The TREC-8 question answering track. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00). European Language Resources Association (ELRA),

Athens, Greece (2000). https://aclanthology.org/L00-1018/

[39] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002) https://doi.org/10.1145/582415.582418

[40] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=SkeHuCVFDr

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SourceQualityAssessmentFramework..pdf
- ExampleSystemOutputsandXAI.pdf