Adaptive Explainable Segmentation of Biomedical Questions for Knowledge Graph-Based Clinical Information Retrieval

Anonymous ACL submission

Abstract

Biomedical question answering (QA) systems often struggle with complex multi-intent queries that mix symptoms, causes and treatments across multiple diseases. We present a segmentation-aware framework that decomposes noisy user input into clean, knowledge graph (KG)-aligned sub-questions. Our pipeline integrates grammatical correction, discourse-aware segmentation, fuzzy disease recognition and a novel SHAP-guided (SHapley Additive exPlanations) embedding fusion mechanism that dynamically weights PubMedBERT, BioBERT and SentenceTransformer models per segment. This enables precise disease aspect pairing and improves interpretability through token-level attribution. Evaluated on a benchmark of straightforward and multi-aspect cardiovascular queries, our method significantly outperforms baseline retrieval and LLM-KG systems in both answer accuracy and semantic alignment with external references. By explicitly structuring user input before KG retrieval, our system improves coverage, factual consistency and transparency.

1 Introduction

Biomedical question answering (QA) has seen major progress with the advent of large language models (LLMs) and improved semantic retrieval techniques (Jin et al., 2022). However, real-world questions posed by clinicians, patients, or researchers often remain challenging. These questions are typically conversational, multi-intent, and contain implicit references that require careful resolution (Gupta, 2012). For example, a user might ask: *"What are the symptoms of heart failure, how is atrial fibrillation treated, and what causes a heart attack?"*. This single query mixes three different diseases and three different aspects (symptoms, treatment, and causes) in one sentence. Without segmentation, such mixed intentions can easily be conflated, leading to incomplete, misleading, or untraceable answers. In practice, most retrieval systems will only focus on the first part of the query or attempt to generate a generic response that overlooks important details. To return accurate and comprehensive answers, the question must first be broken down into clear sub-questions—*"What are the symptoms of heart failure?"*, *"How is atrial fibrillation treated?"*, and *"What causes a heart attack?"*, so that each disease–aspect pair can be precisely matched to the corresponding knowledge graph entry. This motivates the need for a robust segmentation mechanism as the foundation for safe and reliable biomedical QA.

Knowledge graphs (KGs) offer a structured way of organising biomedical knowledge (Ling et al., 2025). By encoding diseases and their attributes as entities and relationships, they enable precise retrieval across fields such as symptoms, causes, diagnosis, and treatment. However, this requires that input questions be reshaped into clear, aspect-specific queries aligned with the KG schema. Biomedical queries often complicate this process through coordination (linking several ideas in one sentence), ellipsis (where parts of the meaning are implied), varied terminology, and ambiguous references. For example, in the question *"What complications can heart failure cause, and what about aortic dissection?"*, the second clause omits both the aspect and the verb, so the system must infer that the user is asking about complications of aortic dissection. Without resolving such ambiguity, the correct disease–aspect pairs cannot be generated for retrieval. These phenomena highlight why segmentation and disambiguation are critical for mapping free-text questions into precise KG queries (Demner Fushman and Friedman, 2024).

Existing segmentation approaches tend to rely only on keyword spotting or semantic similarity. Keyword-based methods are brittle to paraphrase and prone to over-segmentation (Bayomi et al., 2015; Feng et al., 2020), often splitting queries at

the wrong points or failing to recognise equivalent phrasing. Purely semantic approaches, by contrast, depend heavily on embedding similarity and can mis-assign aspects when the disease or entity is not explicitly mentioned (Xiong et al., 2023; Maraj, 2024). In both cases, these limitations lead to either missing important sub-questions or introducing irrelevant ones. Moreover, many systems rely exclusively on a single embedding model, ignoring the variability in how different models capture biomedical phrasing and terminology. This lack of model diversity can result in biased or incomplete retrieval, since some encoders are better tuned for clinical terminology while others are stronger at handling general paraphrasing.

In this work, we introduce a segmentation-aware preprocessing module for KG-based biomedical QA. This module is not a user-facing front end, but rather a core pipeline component designed to transform noisy, multi-intent questions into well-structured inputs that align with the knowledge graph schema. It consists of three key components: a grammar correction model (Section 3.3), a discourse-aware segmentation mechanism (Section 3.4), and an SHAP-based adaptive fusion strategy (Section 3.1). These elements work together to improve the alignment and interpretability of disease–aspect pairs, ensuring that each sub-question maps cleanly to the appropriate KG entry. The segmentation and adaptive fusion mechanisms were specifically developed for this study, while the grammar correction model is adapted from existing work (Li et al., 2022). We evaluate the complete system on a cardiovascular benchmark covering both straightforward and multi-intent queries, demonstrating that it consistently outperforms strong baselines in terms of precision, completeness, and semantic alignment. This paper therefore, emphasises segmentation and disease–aspect linking as a reliable foundation for safe, transparent, and clinically meaningful end-to-end QA grounded in validated knowledge graphs.

2 Related Work

Recent work increasingly combines large language models (LLMs) with knowledge graphs (KGs) to enhance retrieval accuracy and reasoning in question answering systems. One representative example is a customer service QA framework that integrates retrieval-augmented generation (RAG) with a KG constructed from historical support tickets (Xu et al., 2024). In this framework, individual tickets are parsed into hierarchical trees to preserve intra-ticket structure, while inter-ticket connections are established through explicit relational links and semantic similarity edges. Each node is then encoded using E5 embeddings (Wang et al., 2022) and indexed in the Qdrant vector database (Ockerman et al., 2025) to enable efficient dense retrieval. During inference, the LLM extracts entities and intents from a user query, retrieves semantically aligned subgraphs, and translates the request into a Cypher query for KG execution, before generating a natural language response from the retrieved evidence. Empirically, the system achieves a 77.6% improvement in mean reciprocal rank (MRR), a 0.32 BLEU gain, and a 28.6% reduction in issue-resolution time when combining GPT-4 with E5 and Qdrant. Despite these promising results, the approach remains highly domain-specific and lacks explicit mechanisms for verifying the provenance of retrieved evidence, ensuring source transparency, or segmenting long multi-intent queries—limitations that can hinder its application to more complex biomedical or multi-domain settings.

Another notable work introduced a medical question answering pipeline that combines traditional information retrieval with neural ranking, aiming to bridge the gap between symbolic and semantic search (Guo et al., 2022). The system first retrieves coarse candidate answers using Elasticsearch, then refines the ranking through a Siamese network built on pretrained language models, which captures semantic similarity between question–answer pairs more effectively than lexical overlap alone. Named Entity Recognition (NER) and a domain-specific medical knowledge graph (KG) are integrated to align entities across questions and candidate answers, leading to a 2–3% improvement in mean average precision (MAP) and mean reciprocal rank (MRR) over baseline BERT models across four benchmark datasets. Despite these gains, the KG in this system serves primarily as a supporting feature for ranking rather than a core reasoning or answer-generation component. Its reliance on a limited open-source medical KG also restricts scalability and generalisation, making it less effective when handling complex, multi-intent, or previously unseen biomedical queries that require compositional reasoning and evidence grounding.

Another recent study tackled hallucination reduction in biomedical QA by integrating large language models with knowledge graphs through the

LangChain framework (Pusch and Conrad, 2024). The system translates natural-language queries into Cypher, validates them through a three-stage process (syntax, node-type, and relation checking), and then executes the corrected queries on the KG before synthesising an answer. Evaluated on a 50-question biomedical benchmark, GPT-4 Turbo successfully generated 47 correct queries, substantially outperforming LLaMA3-70B, which achieved 23 (Pusch and Conrad, 2024). Prompt optimisation further improved open-source models, underscoring the importance of query verification for factual grounding. However, the approach still depends on exact string matching and lacks similarity-based entity linking, limiting its robustness to linguistic variation. Moreover, it does not yet support end-to-end verification of retrieved evidence, an issue similarly noted in other LLM–KG QA frameworks (Guo et al., 2022).

Another key issue not directly addressed in these studies is question segmentation. In healthcare and other knowledge-intensive domains, users often pose long or compound questions that mix several aspects, such as symptoms, causes, and treatment of a disease in one query. If processed as a single unit, retrieval and ranking models may struggle to find relevant evidence, lowering accuracy (Guo et al., 2022; Pusch and Conrad, 2024). Segmenting such questions into aspect-focused sub-queries improves semantic matching and increases the chance of retrieving complete answers. It also facilitates multi-hop reasoning, since shorter sub-questions can be mapped more easily to one-, two-, or three-hop paths (Pusch and Conrad, 2024). After retrieving answers, the system can recombine them into a coherent response. A robust segmentation module, whether rule-based, embedding-driven, or neural, is therefore essential for hybrid LLM–KG systems to handle long, complex queries effectively.

3 Methods

Following the work of Alshammari and Atapour-Abarghouei (2025), which introduced a cardiovascular question-answering system grounded in a validated knowledge graph (KG), this study extends that framework with several key improvements. The earlier system relied on data from authoritative medical sources, including the Mayo Clinic (Mayo Clinic, 2025), Cleveland Clinic (Cleveland Clinic, 2025), the NHS (National Health Service (NHS), 2025), the British Heart Foundation (British Heart Foundation, 2025), and MedlinePlus (U.S. National Library of Medicine, 2025), all of which were systematically validated before being ingested into the KG. User queries and disease- or aspect-specific KG nodes were encoded using PubMedBERT, BioBERT, and a general-purpose Sentence Transformer, with similarity scores combined through a fixed weighting scheme.

In the prior approach, disease identification was primarily performed through exact string matching, with a semantic similarity fallback, while aspect detection utilised a keyword- and lemma-based matcher supplemented by embeddings. Longer or compound queries were segmented only by basic sentence splitting or conjunction-based rules. For each identified disease–aspect pair, a parameterised Cypher query retrieved the corresponding text, source, references, and images from the KG. If the overall confidence score fell below a threshold, a Text-to-Text Transfer Transformer (T5)–based paraphrasing loop (Raffel et al., 2020) was invoked to reduce mismatches between user phrasing and KG terminology.

Building on this foundation, the present work introduces adaptive encoder weighting, grammar-aware normalisation, multi-layered disease recognition, and discourse-level segmentation. These enhancements are specifically designed to improve robustness, interpretability, and performance, addressing the earlier system's limitations in handling long, compound queries, informal phrasing, and misspelt medical terms. Figure 1 shows an overview of the pipeline.

3.1 Adaptive SHAP-weighted embeddings

We introduce an adaptive SHAP-driven weighting mechanism that dynamically balances PubMedBERT, BioBERT, and Sentence Transformer embeddings. SHAP attributes a predictive score to each input feature by computing its marginal contribution across all feature combinations (Lundberg and Lee, 2017). In our system, token-level importance values are computed separately for each encoder on each question segment, normalised, and then used as dynamic weights to combine similarity scores with disease names, aspects, and disease overviews. This adaptive fusion is both more accurate and more interpretable than applying a static mixture of models to every query. The pipeline highlights which encoder is most informative for each segment. As a result, it becomes clear whether PubMedBERT, BioBERT, or the Sentence Trans-

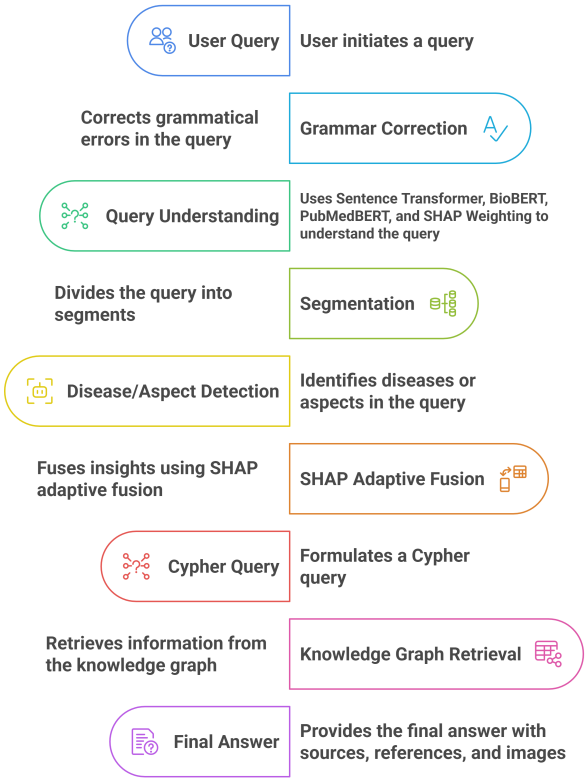

Figure 1: Overview of the proposed segmentation-aware knowledge graph QA pipeline.

former provides the decisive signal. Such transparency is crucial in a medical QA setting, where identifying the representation that drives a match supports error diagnosis, explanation of system behaviour, and assurance that retrieved answers are grounded in the most appropriate biomedical language model for the given question segment.

3.2 Enhanced disease recognition

Rather than relying on exact matches or a single similarity threshold, the new pipeline integrates boundary-aware matching, semantic similarity to cached embeddings, noun-chunk extraction, and fuzzy matching. Fuzzy matching supports approximate string comparison to recover matches despite typographical errors or uncommon variants (Dexter and Benouarets, 1997; Rees, 2014). Candidate diseases are then shortlisted by comparing each segment to pre-computed disease embeddings using the SHAP-weighted ensemble.

This layered design overcomes the brittleness of exact string matching, which struggles with synonyms, abbreviations, and compound names, and the limitations of threshold-based similarity, which often fails with overlapping terminology or multiple conditions in a single query. By combining complementary detection strategies before SHAP-weighted scoring, the pipeline increases recall, improves precision by filtering out spurious matches, and provides transparent attribution of contributing features (see Tables 1 and 3). This results in more reliable disease recognition and cleaner aspect–disease pairs for segmentation and answer generation, improving coverage, factual accuracy, and user trust.

3.3 Grammar correction

A BERT-based sequence-to-sequence grammar correction model (Li et al., 2022) is applied to normalise user input before segmentation. It corrects spelling and grammatical errors so disease and aspect terms are consistently recognised, even when misspelled or irregularly phrased. This standardisation prevents mismatches in detection and reduces segmentation errors, improving downstream reliability.

3.4 Redesigned segmentation for multi-aspect queries

The segmentation module is fundamentally overhauled to cope with the complexity and variability of real-world biomedical questions, particularly those that conflate multiple aspects and diseases in a single utterance. In the earlier pipeline, segmentation relied on a simple clause splitter, which frequently under-segmented long questions (missing key aspects) or over-segmented at inappropriate points (producing incomplete fragments). To overcome these shortcomings, the new module integrates sentence boundary detection (Rehbein et al., 2020) with a cautious, discourse-aware strategy that only inserts splits at conjunctions such as “and” or “or” when both sides show clear aspect-specific cues. This ensures that each resulting segment reflects a coherent and medically meaningful intent rather than an arbitrary slice of text.

Aspect detection is further reinforced by an expanded lexicon of single- and multi-word patterns combined with lemma-based matching (El-Shishtawy and El-Ghannam, 2014), which provides the opportunity for recognition of a far broader range of ways in which clinical aspects (e.g., symptoms, causes, complications, treatments) are expressed (Maraj, 2024). Each detected segment is then analysed to extract explicit diseases and aspects. When segments contain several aspects or multiple diseases, the module decomposes them

into smaller, aspect-focused units and uses a global shortlist derived from the corrected question to guide disease assignment. Where several diseases are plausible for the same segment, a composite scoring function (see Section 3.2) integrates the proximity of each disease mention within the full query with the semantic similarity between the segment and each disease's overview to select the most likely match.

The segmentation process produces stand-alone sub-queries that are canonicalised, meaning that each is normalised into a standard, unambiguous form where exactly one disease entity is paired with exactly one aspect type from the KG schema (e.g., symptoms, causes, complications, treatments). Canonicalisation also eliminates redundant phrasing, resolves implicit references, and rewrites elliptical fragments into full queries.

For example, given the user question: *"What are the complications of heart failure, and what about aortic dissection?"*, the raw segmentation may yield: *"What are the complications of heart failure?"* and *"What about aortic dissection?"*.

After canonicalisation, the elliptical fragment is rewritten into a fully specified form: *"What are the complications of heart failure?"* and *"What are the complications of aortic dissection?"*.

This ensures that all sub-queries conform to the same disease–aspect format, making them directly executable as Cypher queries against the KG.

3.5 Evaluation Setup

To quantify the impact of the proposed enhancements, we constructed a dedicated cardiovascular question answering benchmark consisting of 100 queries in total, 50 straightforward questions focusing on a single disease–aspect intent and 50 complex queries that require segmentation across multiple aspects or diseases. This benchmark is specifically designed to test system behaviour under both simple and challenging conditions and it allows for a direct comparison of our redesigned pipeline with strong baselines adapted from recent related work (Guo et al., 2022; Jin et al., 2022).[1]

System outputs were evaluated using three complementary automatic metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). BLEU measures the precision of overlapping n-grams between a system output and a reference text, ROUGE emphasises recall of overlapping n-grams and longest common subsequences and BERTScore computes similarity between contextual embeddings, which makes it less sensitive to paraphrasing and word-order differences and thus more indicative of true semantic alignment.

Because disease names and relationship/aspect labels are typically short lexical items rather than full sentences, we report BLEU-1 as the primary n-gram precision metric for these targets. This captures exact token matches without penalising brevity and also include ROUGE-1 to measure recall of key tokens (Guo et al., 2022). For answer generation, which yields longer free-text outputs, we report BLEU-1/2/3/4 together with ROUGE-1/2/L to assess both unigram overlap and higher-order n-gram consistency with the reference answers. This combination balances sensitivity to exact medical terminology with the ability to detect more nuanced mismatches in phrasing and content.

In addition to overlap with our KG ground-truth answers, we perform external validation using MedQuAD (Medical Question Answering Dataset) (Ben Abacha and Demner-Fushman, 2019) as a held-out reference. MedQuAD consists of tens of thousands of consumer-health Q&A pairs curated from National Institutes of Health (NIH) websites and is widely used to assess the robustness of medical QA systems to stylistic and phrasing variability. Because MedQuAD does not cover every cardiovascular disease or intent represented in our benchmark, any missing disease–aspect pairs were substituted with reference passages from clinical, public-facing summaries on NHS (National Health Service (NHS), 2025) or Cleveland Clinic (Cleveland Clinic, 2025) websites. These sources were used solely to construct an external target text for scoring and were never incorporated into our KG or retrieval index during evaluation.

The stylistic and structural differences between external references and our KG ground-truth text mean that n-gram metrics such as BLEU and ROUGE can underestimate performance even when the retrieved content is semantically correct. We therefore also report BERTScore (F1), which is less sensitive to paraphrasing and thus a more reliable indicator of semantic alignment. In cases where answer_external BLEU or ROUGE values appear modest, a comparatively higher BERTScore indicates that the system nevertheless produces content that is semantically consistent with the external reference despite differences in wording.

[1] The full benchmark will be publicly released post review to preserve anonymity.

4 Results

We evaluate our pipeline on the benchmark described in Section 3, comparing it with strong baselines. The benchmark covers two regimes: straightforward single-aspect queries and complex multi-aspect or multi-disease queries requiring segmentation.

Baselines performed well on straightforward questions (see Tables 1 and 2), with high disease identification and generally strong aspect detection. However, on multi-aspect or multi-disease queries, they typically detect only the first disease or aspect, returning narrow answers. GPT-4- and GPT-5-based retrieval systems (Pusch and Conrad, 2024; Guo et al., 2022) also show grounding issues, sometimes omitting evidence from the KG or introducing extraneous material. These results highlight the need for explicit segmentation, comprehensive disease–aspect pairing and strict grounding for factual completeness.

In both straightforward and multi-aspect scenarios, our system consistently achieves complete coverage of all disease–aspect pairs and produces fully grounded answers. On straightforward queries (Tables 1 and 2), it reaches perfect disease and relationship matching as well as exact alignment with KG ground truth answers. On complex multi-aspect or multi-disease queries (Tables 3 and 4), it maintains near-perfect scores where baselines drop sharply, confirming that segmentation and layered detection prevent the omission of relevant evidence. Most importantly, external validation against MedQuAD (Ben Abacha and Demner-Fushman, 2019), NHS (National Health Service (NHS), 2025), and Cleveland Clinic (Cleveland Clinic, 2025) demonstrates that the answers are both faithful to the KG and semantically consistent with independent trusted sources. Strong BERTScore values in this setting provide robust evidence of reliability, underscoring that the system's outputs remain clinically accurate and externally validated.

4.1 Disease and Relationship Matching on Straightforward Questions

We first evaluate disease and relationship detection on single-aspect cardiovascular questions (Table 1). Across all baselines, disease identification is consistently strong, with BLEU-1 scores above 0.8 and ROUGE-1 scores above 0.9. For instance, the hybrid Elasticsearch–BERT–NER baseline achieves perfect disease recognition (1.00 BLEU-1 / ROUGE-1), but its aspect matching drops markedly to 0.45 BLEU-1 and 0.64 ROUGE-1, indicating that it often detects the correct disease but fails to capture the intended relationship or aspect type. Similarly, GPT-4 with E5 embeddings (Wang et al., 2022) performs well on disease recognition (0.83 BLEU-1) yet yields inconsistent aspect identification. These findings echo prior observations that purely lexical or embedding-based retrieval pipelines struggle to model aspect-level context (Guo et al., 2022; Pusch and Conrad, 2024). In contrast, our segmentation-aware system achieves perfect precision and recall across both disease and relationship detection (1.00 BLEU-1 / ROUGE-1 for both), showing that it can consistently disambiguate the aspect being queried and align each input with the correct disease–aspect pair. This result highlights that even in simple, single-intent cases, explicit normalisation, segmentation, and SHAP-weighted fusion contribute to stable and fully grounded matches between user queries and KG entries.

Approach	Dis. B1	Dis. R1	Rel. B1	Rel. R1
GPT-4 + E5	0.83	0.83	0.98	0.98
ES + BERT + NER	1.00	1.00	0.45	0.64
GPT-4-Turbo (LC)	1.00	1.00	0.85	0.92
GPT-5 Retrieval	0.96	0.97	0.98	0.99
Ours	**1.00**	**1.00**	**1.00**	**1.00**

Table 1: Disease and relationship matching on straightforward questions. Dis. = Disease, Rel. = Relationship, B1 = BLEU-1 (Papineni et al., 2002), R1 = ROUGE-1 (Lin, 2004).

4.2 Answer Generation on Straightforward Questions

We next assess the quality of full answer generation for single-aspect cardiovascular questions. The quantitative outcomes of this experiment are reported in Table 2. As shown, both GPT-4-Turbo and GPT-5 retrieval pipelines achieve strong n-gram overlap (BLEU-4 = 0.84 and 0.81, respectively) and good semantic similarity (BERTScore = 0.84 and 0.90). However, their outputs frequently omit essential information or provide only partial coverage of the knowledge graph content, occasionally leading to incomplete or clinically ambiguous answers. By contrast, our redesigned pipeline achieves perfect BLEU-1/2/3/4 and ROUGE-1/2/L scores and the highest external BERTScore (1.00 across all metrics), indicating exact lexical alignment with the validated KG and full semantic consistency with independent references. These find-

Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1	External
GPT-4 + E5	0.24	0.13	0.09	0.05	0.35	0.12	0.20	0.61	0.58
ES + BERT + NER	0.42	0.38	0.37	0.36	0.53	0.46	0.49	0.77	0.58
GPT-4-Turbo (LC)	0.84	0.84	0.84	0.84	0.87	0.86	0.87	0.84	0.60
GPT-5 Retrieval	0.81	0.81	0.81	0.81	0.84	0.83	0.83	0.90	0.58
Ours	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.62**

Table 2: Answer generation results on straightforward single-aspect questions. Metrics: BLEU-1/2/3/4 (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), and BERTScore F1 (Zhang et al., 2020). External validation compares system outputs with references from MedQuAD (Ben Abacha and Demner-Fushman, 2019), NHS (National Health Service (NHS), 2025), Cleveland Clinic (Cleveland Clinic, 2025), and Mayo Clinic (Mayo Clinic, 2025).

ings in Table 2 demonstrate that the proposed system produces complete, contextually accurate, and factually grounded answers that more faithfully reflect the underlying medical knowledge.

4.3 Disease and Relationship Matching on Multi-Aspect / Multi-Disease Questions

We then evaluate disease and relationship matching on complex multi-aspect and multi-disease queries that require explicit segmentation (Table 3). In this setting, baseline systems show clear limitations. For example, GPT-4 with the E5 Transformer (Wang et al., 2022) achieves only 0.49 BLEU-1 and 0.56 ROUGE-1 at the disease level, while GPT-5 Retrieval reaches 0.61 BLEU-1 and 0.70 ROUGE-1 for relationship detection. Because these models process the query as a single unit without segmentation, they typically recognise only the first disease or aspect mentioned and overlook others that are explicitly present in the input, resulting in incomplete and narrowly focused matches. By contrast, our pipeline explicitly segments each query into coherent sub-questions before matching, enabling it to capture all disease–aspect mentions expressed in the original query. Consequently, it achieves near-perfect scores (1.00 BLEU-1 / ROUGE-1 for diseases and 0.99 for relationships), demonstrating that segmentation is the key factor enabling comprehensive coverage and accurate disease–aspect alignment in complex biomedical questions.

Approach	Dis. B1	Dis. R1	Rel. B1	Rel. R1
GPT-4 + E5	0.49	0.56	0.89	0.92
ES + BERT + NER	0.74	0.88	0.54	0.61
GPT-4-Turbo (LC)	0.61	0.99	0.83	0.90
GPT-5 Retrieval	0.92	0.92	0.61	0.70
Ours	**1.00**	**1.00**	**0.99**	**0.99**

Table 3: Disease and relationship matching on multi-aspect and multi-disease questions. Dis. = Disease, Rel. = Relationship, B1 = BLEU-1 (Papineni et al., 2002), R1 = ROUGE-1 (Lin, 2004).

4.4 Answer Generation on Multi-Aspect / Multi-Disease Questions

We next examine full answer generation for the multi-aspect and multi-disease portion of the benchmark (Table 4). In this more challenging setting, all baseline systems, including GPT-4-based pipelines and GPT-5-style retrieval, produce responses that are abbreviated, only partially grounded, and primarily focused on the first disease–aspect combination identified within the query. Consequently, key information from subsequent nodes is often omitted, and some outputs include incomplete or extraneous details due to the absence of explicit segmentation or fine-grained evidence control. For instance, GPT-4 with the E5 Transformer (Wang et al., 2022) attains only 0.22 BLEU-1 and 0.04 BLEU-4, while GPT-5 Retrieval achieves 0.37 BLEU-1 and 0.36 BLEU-4, reflecting limited coverage across all disease–aspect pairs expressed in the queries. By contrast, our pipeline explicitly segments each multi-intent question into well-formed sub-queries, ensuring that every detected disease and aspect is individually represented and answered. This segmentation, coupled with SHAP-weighted encoder fusion and layered detection, enables full retrieval of KG evidence across all relevant nodes rather than concentrating on the first identified match. As shown in Table 4, this approach achieves near-perfect scores (BLEU-1/2/3/4 and ROUGE-1/2/L = 0.99; BERTScore = 0.99), confirming that segmentation is the critical factor enabling comprehensive, faithfully grounded, and semantically consistent answers across all disease–aspect combinations.

4.5 Segmentation Performance

We also evaluate the segmentation module independently on a held-out set of annotated questions, comparing predicted sub-queries against gold segmentations using BLEU, ROUGE, and exact-match metrics. This analysis isolates the segmentation

Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1	External Validation
GPT-4 + E5	0.22	0.11	0.07	0.04	0.34	0.10	0.16	0.57	0.55
ES + BERT + NER	0.48	0.44	0.43	0.42	0.60	0.52	0.48	0.74	0.58
GPT-4-Turbo (LC)	0.28	0.28	0.27	0.27	0.33	0.31	0.32	0.55	0.43
GPT-5 Retrieval	0.37	0.37	0.36	0.36	0.42	0.40	0.39	0.60	0.47
Ours	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.62**

Table 4: Answer generation results on multi-aspect and multi-disease questions requiring segmentation. Metrics: BLEU-1/2/3/4 (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), and BERTScore F1 (Zhang et al., 2020). External validation uses references from MedQuAD (Ben Abacha and Demner-Fushman, 2019), NHS (National Health Service (NHS), 2025), Cleveland Clinic (Cleveland Clinic, 2025), and Mayo Clinic (Mayo Clinic, 2025).

component to verify its reliability before integration with disease and aspect detection.

Table 5 summarises the results. The module achieves perfect n-gram and semantic overlap with the gold segments (BLEU-1/2/3/4 and ROUGE-1/2/L all = 1.00), indicating that the generated sub-questions are lexically and semantically identical to the reference segments. At the question level, 96% of examples exactly match the gold segment sets, and an equal proportion match the reference segment count. Even in the few non-exact cases, all gold segments were recovered, with only minimal variations in word order or phrasing that do not affect meaning.

The system correctly segmented every benchmark question except one. For the input *"Discuss the initial and advanced symptoms, as well as the major cardiovascular complications seen in children suffering from Kawasaki disease if not treated in time"*, the model produced three sub-questions—*What are the symptoms of Kawasaki disease?*, *How is Kawasaki disease treated?*, and *What are the complications of Kawasaki disease?*—while the gold segmentation included only two. The misinterpretation of the phrase *"if not treated in time"* as a treatment cue led to a single over-segmentation, as the system interpreted this conditional clause as implying a separate treatment-related intent. This case highlights that although the module handles explicit aspect cues accurately, conditional or implicit phrasing can still create ambiguity in intent boundaries, a challenge common across discourse-based segmentation models.

Overall, these results confirm that the segmentation module can robustly decompose complex, multi-intent biomedical questions into precise, aspect-specific sub-queries. Its high lexical and semantic fidelity ensures that each resulting question aligns cleanly with the KG schema, enabling accurate disease–aspect mapping and preventing information loss during retrieval. Residual errors mainly arise from implicit or conditional phrasing, which future iterations may address through improved discourse cue detection.

Metric	Score
BLEU-1 / 2 / 3 / 4	1.00 / 1.00 / 1.00 / 1.00
ROUGE-1 / 2 / L (F1)	1.00 / 1.00 / 1.00
Exact match rate	96%
Segment count match rate	96%
Average segment overlap ratio	1.00

Table 5: Segmentation performance on the annotated cardiovascular question set.

5 Conclusion

We have presented a segmentation-aware, knowledge-graph-grounded cardiovascular question answering system that combines discourse-aware splitting with SHAP-guided adaptive fusion of multiple biomedical encoders and layered disease recognition (Matsumoto et al., 2024). By explicitly decomposing complex multi-intent queries into well-formed sub-questions and dynamically weighting PubMedBERT, BioBERT, and Sentence Transformer embeddings at the segment level, the system achieves high precision, completeness, and semantic alignment on both straightforward and multi-aspect cardiovascular queries. Our evaluation shows that this approach substantially improves disease and relationship detection, answer generation quality, and segmentation accuracy compared with strong LLM–KG baselines, achieving 100% n-gram overlap and a 96% exact match rate for segmentation. These results demonstrate that carefully designed front-end processing can dramatically enhance the reliability and interpretability of retrieval-augmented biomedical QA. Beyond cardiovascular content, the pipeline is in principle domain-agnostic. Future work will extend coverage to other specialties, test scalability and latency in real-time settings, and incorporate expert-in-the-loop validation to complement automatic metrics.

Limitations

This study is limited to the cardiovascular domain. Although the redesigned pipeline is in principle domain-agnostic, all resources, including the knowledge graph, aspect lexicons and evaluation items, are specific to cardiovascular diseases. The generalisability of the approach to other medical specialities or to heterogeneous, multi-domain KGs has not yet been assessed.

System performance also depends on the coverage and currency of the underlying knowledge graph. Even though the content was curated from authoritative sources and validated before ingestion, any omissions, outdated information or source biases may propagate into retrieval and answer-generation. The present work does not quantify the effect of KG incompleteness nor explore automated update strategies.

Evaluation relies on automatic metrics such as BLEU, ROUGE, and BERTScore computed against the KG ground truth and an external reference. While BERTScore can capture deeper semantic similarity between system outputs and references, n-gram–based metrics such as BLEU and ROUGE are limited to surface lexical overlap and cannot reliably reflect semantic correctness. This mismatch means that clinically accurate paraphrases may be undervalued by n-gram metrics, while semantically incomplete responses may still receive artificially high scores from embedding-based measures. Prior work has highlighted these shortcomings, showing that BLEU in particular lacks validity outside machine translation and can mis-rank systems under certain conditions (Reiter, 2018; Mathur et al., 2020). These limitations underscore the need for complementary human validation to ensure that evaluation outcomes truly align with clinical adequacy and user expectations.

Segmentation may still under-segment when aspect cues are implicit or over-segment in the presence of coordination ambiguities. Fuzzy matching improves recall for misspellings and variants but can increase false positives for short or polysemous disease strings. Grammar correction can over-normalise rare clinical terms. SHAP-guided adaptive weighting improves alignment but adds computational overhead and requires careful normalisation to avoid segment-level instability.

Ethical Considerations

This system uses only publicly available and trusted medical sources such as the Mayo Clinic (Mayo Clinic, 2025), Cleveland Clinic (Cleveland Clinic, 2025), NHS (National Health Service (NHS), 2025), and MedlinePlus (U.S. National Library of Medicine, 2025), and does not use any private or personal health data. All content incorporated into the knowledge graph was systematically validated prior to ingestion using a structured reliability assessment that considered accuracy, authority, currency, and purpose of each source. This process was designed to minimise the risk of misinformation and ensure that retrieved answers remain aligned with authoritative medical knowledge.

Although the system is designed to improve reliability and interpretability in medical question answering, it is not a diagnostic tool and is not intended to replace the judgement of qualified healthcare professionals. Users are explicitly advised to consult medical practitioners for personalised advice or urgent medical concerns.

The segmentation module enforces that multi-intent queries are decomposed into clear, clinically meaningful sub-questions, which reduces the risk of users receiving incomplete or conflated answers. Similarly, the SHAP-guided adaptive weighting mechanism increases transparency by showing which encoder contributes most to each decision, thereby making potential biases in retrieval more visible and traceable. These measures support the goal of delivering accurate, evidence-based, and explainable answers, while reducing the risk of hallucination or misleading outputs.

References

Mohammed Alshammari and Amir Atapour-Abarghouei. 2025. Reliable cardiovascular question answering with validated knowledge graphs and multi-model embeddings. Preprint.

Mostafa Bayomi, Killian Levacher, M. Rami Ghorab, and Séamus Lawless. 2015. Ontoseg: A novel approach to text segmentation using ontological similarity. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1274–1283. IEEE.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):954–959.

British Heart Foundation. 2025. British heart foundation – heart and circulatory diseases. https://www.

bhf.org.uk/informationsupport/conditions. Accessed: 2025-10-02.

Cleveland Clinic. 2025. Cleveland clinic – health library. https://my.clevelandclinic.org/health. Accessed: 2025-10-02.

Dina Demner Fushman and Carol Friedman. 2024. Overview of linguistic information. In Hua Xu and Dina Demner Fushman, editors, *Natural Language Processing in Biomedicine: A Practical Guide*, pages 13–42. Springer, Cham.

A. L. Dexter and M. Benouarets. 1997. Model-based fault diagnosis using fuzzy matching. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 27(5):673–682.

Tarek El-Shishtawy and Fatma El-Ghannam. 2014. A lemma based evaluator for semitic language text summarization systems. *Preprint*, arXiv:1403.5596.

L. Feng, Y. Niu, Z. Liu, J. Wang, and K. Zhang. 2020. Discovering technology opportunity by keyword-based patent analysis: A hybrid approach of morphology analysis and usit. *Sustainability*, 12(1):136.

Q. Guo, S. Cao, and Z. Yi. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564.

P. Gupta. 2012. A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4).

Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, X. Liu, S. Yu, C. Tan, M. Chen, and S. Huang. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*.

J. Li, J. Guo, Y. Zhu, X. Sheng, D. Jiang, B. Ren, and L. Xu. 2022. Sequence-to-action: Grammatical error correction with action guided sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10974–10982.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pages 74–81.

I. Ling, Y. Wu, G. Su, and X. Li. 2025. Zero-shot construction of chinese medical knowledge graph with GPT-3.5-Turbo and GPT-4. *ACM Transactions on Management Information Systems*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30.

A. Maraj. 2024. Contextual topics: advancing text segmentation through pre-trained models and contextual keywords. Master's thesis / repository.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btae353.

Mayo Clinic. 2025. Mayo clinic – medical information and health advice. https://www.mayoclinic.org/. Accessed: 2025-10-02.

National Health Service (NHS). 2025. Nhs – conditions and treatments. https://www.nhs.uk/conditions/. Accessed: 2025-10-02.

Seth Ockerman, Amal Gueroudji, Song Young Oh, Robert Underwood, Nicholas Chia, Kyle Chard, Robert Ross, and Shivaram Venkataraman. 2025. Exploring distributed vector databases performance on hpc platforms: A study with qdrant. *Preprint*, arXiv:2509.12384.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Lukas Pusch and Tobias O. F. Conrad. 2024. Combining llms and knowledge graphs to reduce hallucinations in question answering. *Preprint*, arXiv:2409.04181.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tony Rees. 2014. Taxamatch, an algorithm for near ('Fuzzy') matching of scientific names in taxonomic databases. *PLOS ONE*, 9(9):e107510.

Ines Rehbein, Josef Ruppenhofer, and Thomas C. Schmidt. 2020. Improving sentence boundary detection for spoken language transcripts. In *International Conference on Language Resources and Evaluation*.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

U.S. National Library of Medicine. 2025. Medlineplus – trusted health information. https://medlineplus.gov/. Accessed: 2025-10-02.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Ying Xiong, Xin Yang, Linjing Liu, Ka-Chun Wong, Qingcai Chen, Yang Xiang, and Buzhou Tang. 2023. EARA: Improving biomedical semantic textual similarity with entity-aligned attention and retrieval augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8760–8771, Singapore. Association for Computational Linguistics.

Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Question	Predicted Segmentation	Gold-standard Segmentation
Discuss the initial and advanced symptoms, as well as the major cardiovascular complications seen in children suffering from Kawasaki disease if not treated in time.	• What are the symptoms of Kawasaki disease? • How is Kawasaki disease treated? • What are the complications of Kawasaki disease?	• What are the symptoms of Kawasaki disease? • What are the complications of Kawasaki disease?
What pathophysiological mechanisms contribute to the causes of heart failure and heart attack?	• What are the causes of heart failure? • What are the causes of heart attack?	• What are the causes of heart failure? • What are the causes of heart attack?
What are the primary causes and contributing factors that can lead to the development of aortic dissection, how do they interact in triggering this serious cardiovascular condition, and when should an individual seek immediate medical attention?	• What are the causes of aortic dissection? • When should someone seek medical attention for aortic dissection?	• What are the causes of aortic dissection? • When should someone seek medical attention for aortic dissection?
What integrated prevention strategies reduce the risk of both heart attack and heart failure, and how do they impact future complications?	• How can heart failure be prevented? • How can heart attack be prevented? • What are the complications of heart failure? • What are the complications of heart attack?	• How can heart failure be prevented? • How can heart attack be prevented? • What are the complications of heart failure? • What are the complications of heart attack?

Table 6: Examples of predicted and gold-standard question segmentations produced by the redesigned system.

Appendix A: Segmentation Results

This appendix illustrates how the redesigned segmentation module decomposes complex, multi-aspect cardiovascular questions into clear sub-questions aligned with the knowledge graph schema. Table 6 presents representative examples comparing predicted segmentations against gold-standard references. In nearly all cases, the system correctly extracts complete aspect–disease pairs. The only notable error occurs for the Kawasaki disease example, where the phrase “if not treated in time” was misinterpreted as a treatment query, leading to an extra sub-question being generated.