QIME: Constructing Interpretable Medical Text Embeddings via Ontology-Grounded Questions

Anonymous ACL submission

Abstract

While dense biomedical embeddings achieve strong performance, their black-box nature limits their utility in clinical decision-making. Recent question-based interpretable embeddings represent text as binary answers to natural-language questions, but these approaches often rely on heuristic or surface-level contrastive signals and overlook specialized domain knowledge. We propose **QIME**, an ontology-grounded framework for constructing interpretable medical text embeddings in which each dimension corresponds to a clinically meaningful yes/no question. By conditioning on cluster-specific medical concept signatures, QIME generates semantically atomic questions that capture fine-grained distinctions in biomedical text. Furthermore, QIME supports a training-free embedding construction strategy that eliminates per-question classifier training while further improving performance. Experiments across biomedical semantic similarity, clustering, and retrieval benchmarks show that QIME consistently outperforms prior interpretable embedding methods and substantially narrows the gap to strong black-box biomedical encoders, while providing concise and clinically informative explanations. Code will be released upon publication.

1 Introduction

The deployment of AI systems in high-stakes biomedical applications requires representations that are not only effective but also human-auditable. Recent advances in dense neural encoders (Devlin et al., 2019; Vera et al., 2025), particularly large pretrained language models, have led to substantial performance gains across biomedical NLP tasks. However, these dense embeddings remain inherently opaque: individual dimensions lack explicit semantic meaning. This lack of transparency hinders error analysis and clinical auditing.

To address this issue, a growing body of work has explored interpretable text embeddings that associate embedding dimensions with human-understandable semantics. As reviewed in Section 2, early efforts include Concept Bottleneck Models (CBMs) (Koh et al., 2020), which introduce predefined concepts as intermediate representations. Anchor-based methods (Wang et al., 2025) represent texts via similarity to reference documents, but interpretation requires inspecting heterogeneous anchor texts, imposing high cognitive burden. More recently, question-based embeddings (Sun et al., 2025; Benara et al., 2024) have emerged, where each dimension corresponds to the answer to a natural-language question. While this paradigm offers more explicit semantics, it suffers from two key limitations in the medical domain: questions are predefined or generated solely using corpus-driven signals that capture surface-level patterns rather than clinically meaningful concepts, and embedding construction incurs substantial computational overhead, either through extensive LLM queries or the training of a large number of supervised classifiers.

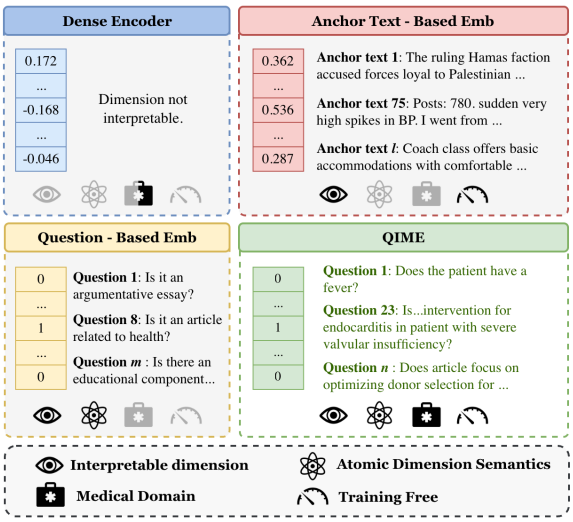

Figure 1: Comparing existing text embedding with the proposed framework.

Therefore, this work is motivated by three observations. First, medical ontologies encode rich and structured domain knowledge that can guide the discovery of clinically meaningful semantic dimensions. Second, the number, granularity, and semantic clarity of the generated questions or anchor texts critically affect the interpretability of the resulting embeddings. Third, practical deployment requires embedding construction that avoids large-scale supervision, costly annotation, or expensive inference-time reliance on large language models (LLM). Figure1 illustrates the distinctions among different embedding paradigms.

To address these challenges, we introduce **QIME**, a framework for constructing **Q**uestion-based **I**nterpretable **M**edical **E**mbeddings grounded in medical ontologies. QIME bridges structured medical knowledge and interpretable natural-language representations through an ontology-grounded question generation process. Specifically, we cluster a large medical corpus and extract biomedical concept signatures for each cluster, which are used to constrain an LLM to generate discriminative, domain-specific questions. In a second stage, QIME constructs sparse, interpretable embeddings based on these questions. Besides classifier-based inference, we further propose a training-free embedding construction strategy based on similarity-driven top-k selection, optionally enhanced with diversity-aware dimension selection via Maximal Marginal Relevance (MMR).

We evaluate QIME on a diverse set of biomedical benchmarks spanning semantic textual similarity, clustering, and information retrieval. Experimental results show that QIME consistently outperforms prior interpretable embedding methods and substantially narrows the performance gap to strong black-box biomedical encoders, while providing explicit and clinically grounded dimensions. Qualitative analyses further demonstrate that QIME produces semantically atomic and clinically informative representations, enabling transparent inspection of model behavior for downstream tasks.

Our contributions are summarized as follows:

- We propose **QIME**, an **ontology-grounded** framework for constructing **question-based interpretable medical** text embeddings, yielding clinically meaningful and discriminative dimensions.
- We develop a **training-free**, **sparse** embedding construction strategy with optional diversity-aware selection, eliminating the need for expensive QA supervision.
- We demonstrate that QIME achieves **strong empirical performance and interpretability** across multiple biomedical similarity, clustering, and retrieval tasks.

2 Related Work

2.1 Black-Box Text Embeddings

Dense neural embeddings dominate modern NLP pipelines for semantic similarity, clustering, and retrieval. Contextual encoders such as BERT (Devlin et al., 2019) and contrastive sentence models like SimCSE (Gao et al., 2021) achieve strong performance. Recent work shows that decoder-only language models can also be repurposed as embedding models (BehnamGhader et al., 2024; Vera et al., 2025).

In the biomedical domain, continued pretraining yields specialized encoders, including BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2022). Ontology-aware models, SapBERT (Liu et al., 2021) and BioLORD (Remy et al., 2024), leverage UMLS synonym sets to improve biomedical representations. Despite their effectiveness, these models produce dense representations whose dimensions lack explicit semantic meaning.

2.2 Interpretable Text Embeddings

To address the opacity of dense encoders, prior work has explored interpretable representations (Opitz et al., 2025). Concept-based approaches, such as CBMs (Koh et al., 2020), TCAV (Kim et al., 2018), and BIERs (Garcia-Olano et al., 2021), rely on predefined or weakly supervised concepts and offer limited flexibility.

A more recent direction is question-based embeddings, where each dimension corresponds to the answer to a binary yes/no question. QA-Emb (Benara et al., 2024) uses LLM prompting to generate interpretable features, but requires querying the LLM for all dimensions at inference time. CQG-MBQA (Sun et al., 2025) introduces Contrastive Question Generation to produce discriminative questions from semantic clusters, reducing inference-time LLM usage by training a classifier for each dimension, at the cost of additional annotation and training overhead. Anchor-based methods such as LDIR (Wang et al., 2025) represent texts via similarity to reference anchors, achieving compact representations but requiring users to interpret

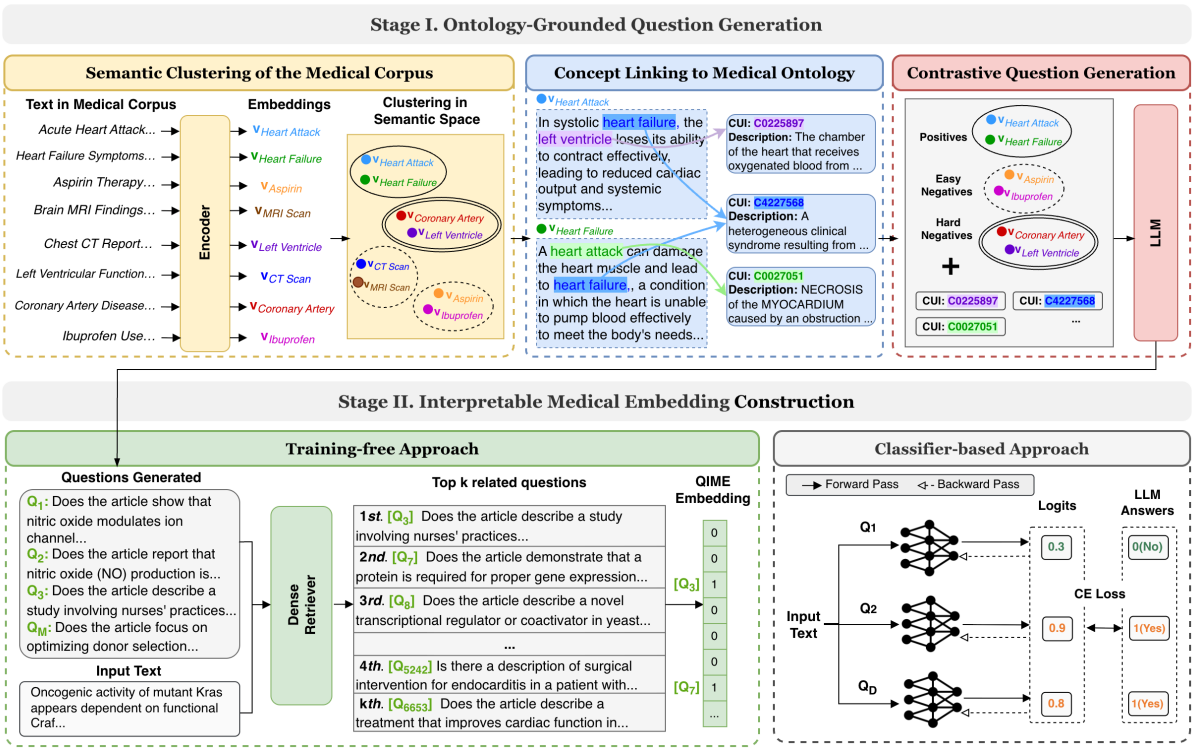

Figure 2: Overview of the QIME framework.

dimensions through long, heterogeneous anchor texts rather than self-describing semantic units.

In contrast to prior methods, QIME grounds question generation in a medical ontology, producing semantically atomic and clinically meaningful dimensions, and constructs sparse embeddings using a training-free, diversity-aware dimension activation strategy.

3 The QIME Framework

3.1 Overview

We propose **QIME** (Ontology-Grounded **Q**uestion-based **I**nterpretable **M**edical **E**mbeddings), a framework for constructing interpretable embeddings for medical text, in which each dimension corresponds to a clinically meaningful natural-language question. QIME aims to produce representations that are both effective for downstream tasks and faithful to medical domain knowledge.

At a high level, QIME represents each document as a sparse binary vector indexed by yes/no medical questions (e.g., *"Does the text describe adverse drug reactions?"*). Unlike prior question-based approaches, QIME does not rely on predefined or heuristic questions; instead, questions are automatically discovered through a contrastive generation process explicitly grounded in medical ontologies and guided by corpus-level structure.

As illustrated in Figure 2, the QIME framework consists of two key stages: (1) **Ontology-grounded question generation**, which discovers clinically meaningful question dimensions from an unlabeled medical corpus, grounded by domain ontology; and (2) **Interpretable embedding construction**, which encodes new texts into sparse question-indexed representations, without requiring QA supervision or classifier training.

3.2 Task Formulation

We now formally define the embedding task addressed by QIME. Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote a large medical text corpus, where each x_i is a document, clinical note, or medical passage. Our goal is to learn an embedding function

$$f : x \mapsto \mathbf{z} \in \{0,1\}^M,$$

here each dimension z_j corresponds to a clinically meaningful yes/no question q_j, and $z_j = 1$ indicates that question q_j is highly relevant to x.

3.3 Ontology-Grounded Question Generation

The objective of the first stage is to discover a set of questions that are both discriminative with respect to the corpus and grounded in clinically

meaningful concepts. Purely data-driven question generation often produces surface-level or stylistic distinctions, which are inadequate for medical interpretation. QIME addresses this issue by combining corpus-level semantic structure with explicit ontology grounding.

Semantic Clustering of the Medical Corpus. We begin by organizing the corpus into semantically coherent regions. Each document x_i is encoded into a dense representation $\mathbf{h}_i$ using a pretrained medical text encoder. We then apply unsupervised clustering to partition the corpus into K clusters, $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{C}_k$, where each cluster $\mathcal{C}_k$ groups texts that are distributionally similar and typically represent a latent medical topic or concept region, such as diagnosis, treatments, or medications. Operating at the cluster level enables question discovery to focus on shared semantic properties across multiple documents, rather than idiosyncratic details of individual instances.

Cluster-Level Ontology Grounding. To align the discovered semantic clusters with established domain knowledge, we ground each cluster in a medical ontology.

For a given cluster $\mathcal{C}_k$, we apply named entity recognition and entity linking to all documents in the cluster to identify medical entities, which are then mapped to ontology concepts. Specifically, we use Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) (Bodenreider, 2004), where each CUI represents a canonical medical concept that unifies synonymous terms across different medical vocabularies. The CUIs extracted from cluster $\mathcal{C}_k$ are aggregated to form a cluster-level concept signature $\mathcal{U}_k = \{u_1, u_2, \ldots, u_{|\mathcal{U}_k|}\}$.

This concept signature provides an explicit representation of the medical semantics associated with the cluster, serving as a domain context for the subsequent question generation.

Grounded Contrastive Question Generation. Given a target cluster $\mathcal{C}_k$ and its concept signature $\mathcal{U}_k$, QIME generates a set of binary medical questions that capture the defining semantic properties of the cluster. We adopt a contrastive question generation (CQG) paradigm (Sun et al., 2025) and enhance it with explicit ontology grounding to ensure medical relevance.

Specifically, for each cluster $\mathcal{C}_k$, we construct three types of examples:

- **Positive samples** Documents drawn from $\mathcal{C}_k$.
- **Hard negatives** Documents from clusters that are semantically proximate to $\mathcal{C}_k$.
- **Easy negatives** Documents from semantically distant clusters.

An LLM is prompted to generate yes/no questions that distinguish positive samples from both hard and easy negatives, while being explicitly conditioned on the ontology concepts in $\mathcal{U}_k$, including concept names and descriptions. By jointly leveraging contrastive supervision and ontology constraints, the generated questions are encouraged to reflect clinically meaningful distinctions that are discriminative at the corpus level rather than superficial lexical differences. Generated questions are aggregated and post-processed to remove low-quality, ambiguous, and redundant entries, resulting in a set of M questions, $\mathcal{Q} = \{q_1, \ldots, q_M\}$. Prompts are provided in Appendix A and post-process details are provided in Appendix B.

3.4 Interpretable Medical Embedding Construction

Once the question set $\mathcal{Q}$ is obtained, the second stage constructs interpretable embeddings for individual documents. We first present a classifier-based approach, and then introduce a training-free alternative that improves scalability.

Classifier-based Embedding Construction. An intuitive approach to constructing question-based interpretable embeddings is to treat each question q_j as a binary prediction task. Given a document x, the embedding value for dimension j can be obtained either by directly querying a large language model to answer q_j with a yes/no response, or by training a separate binary classifier for each question using annotated questionanswer pairs. The classifier-based formulation reduces reliance on LLMs at inference time. We provide details of the classifier training procedure in Appendix C.

Training-Free Sparse Embedding Construction. To address the computational and annotation overhead associated with LLM-based inference and per-question classifier training, QIME proposes a training-free embedding construction strategy, referred to as QIME-TF. This variant instantiates QIME using similarity-based question selection without requiring supervised question–answer labels or classifier training.

Type	Model	Clustering (V-Measure ↑)						STS (SC ↑)
		BioP2P	BioS2S	MedP2P	MedS2S	ClusTREC	Average	BIOSSES
Black-Box	**BERT**	29.95	24.40	26.13	23.63	74.50	35.72	54.70
	GloVe	29.32	18.74	26.14	20.49	74.15	33.77	44.93
	SimCSE (Unsup)	30.10	22.94	28.03	25.62	76.41	36.62	68.86
	SimCSE (Sup)	31.91	25.70	28.38	25.85	76.54	37.68	67.19
	MedEmbed	**40.10**	**35.99**	**33.12**	30.44	**83.26**	**44.58**	86.99
	EmbeddingGemma	36.95	33.06	31.68	**30.45**	82.57	42.94	80.46
	PubMedBERT	34.37	30.97	32.36	28.12	82.59	41.68	83.96
	BioLORD	31.30	27.87	31.77	30.28	80.03	40.25	**87.18**
	SapBERT	31.00	20.53	29.43	22.86	77.05	36.17	82.48
	MedCPT	35.11	32.74	30.49	29.29	77.77	41.08	81.95
	BMRetriever	34.48	20.34	29.81	22.62	79.39	37.33	68.85
Interpretable	**Bag-of-Words**	4.73	3.32	12.43	13.05	65.68	19.84	68.78
	LDIR-500	32.39	29.36	30.00	28.98	79.54	40.05	79.30
	CQG-MBQA	34.88	31.14	31.02	28.65	79.67	41.07	54.97
	QA-Emb	24.60	21.11	25.53	22.82	75.30	33.87	46.43
	QIME	38.18	34.82	33.61	**32.00**	79.43	43.61	61.88
	QIME-TF	40.26	**36.83**	33.78	31.83	81.69	44.88	75.60
	QIME-TF-MMR	**40.37**	36.78	**33.92**	31.44	**81.99**	**44.90**	**79.66**

Table 1: Clustering performance measured by V-Measure and semantic textual similarity (STS) measured by Spearman Correlation (SC) across biomedical benchmarks.

Given a document x, we encode it into a dense vector $\mathbf{h}(x)$ and similarly encode all questions into $\{\mathbf{h}(q_j)\}_{j=1}^{M}$ using MedEmbed.

We then compute cosine similarity $s_j = \mathrm{sim}(\mathbf{h}(x), \mathbf{h}(q_j))$, and activate only the top-k most relevant question dimensions:

$$z_j = \begin{cases} 1, & \text{if } q_j \in \text{Top-}k(s_1, \ldots, s_M), \\ 0, & \text{otherwise.} \end{cases}$$

While this relevance-based selection captures the most salient dimensions for each document, similar questions may still introduce redundancy among the activated dimensions. To address this, we further introduce a diversity-aware variant, QIME-TF-MMR, which incorporates maximal marginal relevance (MMR) during top-k selection. Specifically, for each instance, questions are selected iteratively by jointly maximizing relevance to the document and dissimilarity to previously selected questions, encouraging the activated dimensions to cover complementary semantic aspects.

Both training-free variants leverage the empirical sparsity of question-based interpretable embeddings, where only a small subset of dimensions is relevant for any given document. By restricting representations to a small, diverse set of activated questions, QIME produces concise and interpretable embeddings.

4 Experiments

4.1 Experimental Setup

Tasks and Datasets. We evaluate interpretable embeddings on three embedding-centric medical NLP tasks: (i) text clustering, (ii) semantic textual similarity (STS), and (iii) information retrieval. These tasks jointly assess topical structure discovery, fine-grained semantic alignment, and query–document matching.

For **clustering**, we use biomedical subsets from the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023), including BiorxivClusteringP2P (BioP2P), BiorxivClusteringS2S (BioS2S), MedrxivClusteringP2P (MedP2P), and MedrxivClusteringS2S (MedS2S). These benchmarks require grouping biomedical preprints based on either titles (S2S) or abstracts (P2P). We additionally evaluate ClusTREC-Covid (ClusTREC) (Katz et al., 2024), a COVID-focused clustering benchmark derived from TREC-COVID literature. We report V-measure for all clustering tasks.

For **STS**, we use BIOSSES (Muennighoff et al., 2023), which contains 100 biomedical sentence pairs annotated for semantic relatedness on a 0–4 scale. We report Spearman correlation for STS.

For **retrieval**, we evaluate NFCorpus (Boteva et al., 2016) and TRECCOVID (COVID) (Voorhees et al., 2020). We further include medical QA retrieval benchmarks PublicHealthQA (PHQA) (Enevoldsen et al., 2025) and MedicalQARe-

Type	Model	Retrieval (nDCG@10 ↑)						
		NFCorpus	PHQA	MedQA	COVID	R2-IYI	R2-PMC	Average
Black-Box	**BERT**	4.30	46.20	9.78	14.78	6.90	1.80	13.96
	GloVe	13.87	62.57	19.95	36.22	7.88	7.31	24.63
	SimCSE (Unsup)	9.88	61.07	24.51	32.71	10.07	6.43	24.11
	SimCSE (Sup)	12.42	65.89	24.27	30.83	8.28	4.94	24.44
	EmbeddingGemma	31.42	78.70	60.05	50.36	12.66	9.20	40.40
	MedEmbed	**37.07**	**82.37**	**74.82**	**75.73**	**14.96**	11.25	**49.37**
	PubMedBERT	26.60	68.42	58.01	44.76	12.77	**12.51**	37.18
	BioLORD	25.49	74.77	61.49	54.89	12.22	6.09	39.16
	SapBERT	26.77	57.38	58.45	33.40	9.27	5.48	31.79
	MedCPT	28.43	53.97	40.46	54.66	6.09	8.04	31.94
	BMRetriever	3.04	38.62	10.15	18.64	11.22	10.79	15.41
Interpretable	**Bag-of-Words**	21.59	42.29	26.01	19.23	6.83	4.86	20.14
	LDIR-500	**27.08**	70.68	**65.69**	47.04	**13.39**	**10.87**	39.13
	CQG-MBQA	9.74	62.27	40.45	28.49	10.58	6.88	26.40
	QA-Emb	3.87	44.95	21.71	22.42	9.10	5.11	17.86
	QIME	15.74	61.74	54.66	46.31	8.94	5.53	32.15
	QIME-TF	21.29	75.04	57.79	57.96	10.09	5.99	38.03
	QIME-TF-MMR	25.09	**75.64**	62.36	**64.65**	11.79	7.08	**41.10**

Table 2: Retrieval performance measured by nDCG@10 across biomedical information retrieval benchmarks.

trieval (MedQA) (Abacha and Demner-Fushman, 2019), as well as reasoning-intensive clinical retrieval from R2MED (Li et al., 2025), using its MTEB variants R2MEDIIYiClinicalRetrieval (R2-IYI) and R2MEDPMCClinicalRetrieval (R2-PMC). nDCG@10 is reported for all datasets.

Baselines. We compare QIME against strong black-box and interpretable baselines.

For **black-box dense encoders**, we include general-domain models BERT (Devlin et al., 2019), GloVe (Pennington et al., 2014), supervised and unsupervised SimCSE (Gao et al., 2021), and the decoder-based embedding model EmbeddingGemma (Vera et al., 2025). To assess domain-specific performance, we evaluate biomedical encoders PubMedBERT (Gu et al., 2022), SapBERT (Liu et al., 2021), and BioLORD-2023 (Remy et al., 2024), as well as retrieval-oriented models MedCPT (Jin et al., 2023), BMRetriever (Xu et al., 2024), and MedEmbed (Balachandran, 2024).

For **interpretable embeddings**, we include a bag-of-words baseline with classical term weighting (Salton and Buckley, 1988), question-based embeddings QAEmb-MBQA (Benara et al., 2024) and CQG-MBQA (Sun et al., 2025), as well as LDIR-500 (Wang et al., 2025), which represents texts via relative similarities to a fixed set of 500 diverse anchor texts selected.

Implementation Details. We preprocess the PubMed corpus (Roberts, 2001) by filtering low-quality and duplicated entries, yielding approximately 25 million paragraphs. We randomly sample 5 million paragraphs (average length 296 tokens) for semantic clustering. Paragraph embeddings are obtained using MedEmbed (Balachandran, 2024), followed by k-means clustering with 2,500 clusters. Medical entity extraction is performed using HunFlair (Weber et al., 2021). We use Qwen3-30B (Yang et al., 2025) as the LLM backbone for grounded question generation. After post-processing, 8,855 questions are retained for embedding construction. The parameter λ in MMR is set to 0.7 for QIME-TF-MMR. The value of k in top-k dimensions is set to 256.

4.2 Main Results

Clustering. Table 1 summarizes performance on clustering and STS benchmarks. Among black-box encoders, domain-specialized models such as MedEmbed, EmbeddingGemma, and BioLORD achieve the strongest overall results, reflecting the benefits of large-scale biomedical pretraining and task-specific optimization. These models provide strong representation quality but offer limited transparency in their embedding dimensions.

Within the interpretable category, QIME consistently outperforms interpretable embedding baselines across clustering tasks. In particular, QIME substantially improves over QA-Emb and CQG-MBQA on STS and all clustering benchmarks, indicating that ontology-grounded question discovery yields more coherent and discriminative semantic representations than manual crafted or purely data-

Input: *The patient, with a history of lung cancer, presented with chest pain; initial concern for heart attack was raised, but a contrast-enhanced CT scan revealed no coronary occlusion. It showed metastatic disease involving the mediastinum.*

Model	Score	Content of the top-rated Dimensions
LDIR-500	0.532	I had to get a chest recently for severe chest pains. My period was due one week from ... Two weeks later, I was a week late with my period and I had a positive pregnancy test.
	0.521	Thickening of the pleural membranes is not a condition which is treatable. It is a symptom of a disease such as asbestosis, treatment is more focused on the underlying cause of the thickening.
	0.497	Cancer June 21 – July 22. People bearing the Cancer sign are so loving, you can almost consider them emotional. Cancers make up the greater part of caring folks on this earth.
CQG-MBQA	0.998	Does the article discuss medical conditions or treatments?
	0.996	Is the article devoid of promoting extreme diets or quick fixes?
	0.996	Does the article provide factual information rather than anecdotal stories?
QIME	0.719	Is there a focus on pain control in terminally ill or cancer patients?
	0.709	Does the article involve the use of computed tomography (CT) for diagnosing cardiovascular conditions in humans?
	0.700	Is there evidence of a bacterial infection affecting heart valves or cardiac tissue?

Table 3: Qualitative comparison of top-ranked embedding dimensions for the same input. Scores correspond to cosine similarity for LDIR-500, classifier logits for CQG-MBQA, and MMR-based relevance scores for QIME.

driven question generation. Compared to LDIR-500, which relies on similarity to anchor texts, QIME achieves higher average clustering performance while providing self-describing question-based dimensions.

The training-free variants further enhance performance. QIME-TF achieves a higher average clustering score compared to QIME, demonstrating that similarity-based top-k activation can effectively replace supervised per-question classifiers. Incorporating maximal marginal relevance during top-k selection (QIME-TF-MMR) yields additional gains on several clustering benchmarks and achieves the strongest overall clustering performance, even surpassing the performance of black-box biomedical encoders.

STS. On BIOSSES, QIME-TF-MMR also substantially outperforms other interpretable embeddings, narrowing the gap to strong black-box biomedical models while maintaining sparse, human-interpretable representations.

Retrieval. Table 2 reports retrieval performance measured by nDCG@10 across a diverse set of biomedical information retrieval benchmarks. Black-box medical encoders, particularly MedEmbed, achieve the best overall retrieval performance, benefiting from large-scale supervision and retrieval-oriented training objectives.

Among interpretable methods, QIME-TF-MMR achieves the strongest average retrieval performance. It attains competitive results on challenging benchmarks such as PHQA, MedQA, and TREC–COVID. These results indicate that ontology-grounded questions, combined with diversity-aware top-k selection, can effectively support query–document matching despite the use of sparse, binary representations.

Overall, while black-box models remain superior in absolute performance, QIME substantially reduces the performance gap between interpretable and dense embeddings, achieving a favorable trade-off between effectiveness and interpretability across a wide range of biomedical tasks.

4.3 Case Study: Interpreting Question-Based Representations

Table 3 compares the top-ranked embedding dimensions produced by different interpretable methods for the same clinical input involving chest pain in a lung cancer patient, where myocardial infarction was ruled out by contrast-enhanced CT and mediastinal metastasis was identified.

For LDIR-500, the highest-scoring dimensions correspond to long anchor texts, including personal anecdotes and non-medical content, providing limited direct insight into which clinical factors drive the representation. CQG-MBQA produces question-based dimensions, but the top-ranked questions are largely generic and fail to capture clinically specific distinctions. In contrast, QIME activates a relatively small set of semantically atomic, medically grounded questions that directly reflect salient aspects of the input, such as CT-based cardiovascular diagnosis and cancer-related pathology. This example highlights how ontology-grounded question generation yields more precise and clinically informative interpretations.

Model	Med G.	Clustering (V-Measure ↑)						STS (SC ↑)
		BioP2P	BioS2S	MedP2P	MedS2S	Covid	Average	BIOSSES
QIME	✓	38.18	34.82	33.61	32.00	79.43	43.61	61.88
	✗	37.29	34.01	32.91	31.12	79.22	42.91	57.77
QIME-TF-MMR	✓	40.36	36.78	33.92	31.44	81.99	44.90	79.66
	✗	38.75	35.60	33.36	31.21	81.21	44.03	74.32
		Retrieval (nDCG@10 ↑)						
		NFCorpus	PHQA	MedQA	COVID	R2-IYI	R2-PMC	Average
QIME	✓	15.74	61.74	54.66	46.31	8.94	5.53	32.15
	✗	15.35	61.07	51.45	44.56	8.82	4.39	30.94
QIME-TF-MMR	✓	25.09	75.64	62.36	64.65	11.79	7.08	41.10
	✗	20.76	74.13	58.70	54.29	11.00	7.27	37.69

Table 4: Ablation results for QIME with and without medical knowledge grounding (Med G.) on biomedical clustering (V-Measure), STS (Spearman correlation), and retrieval tasks (nDCG@10).

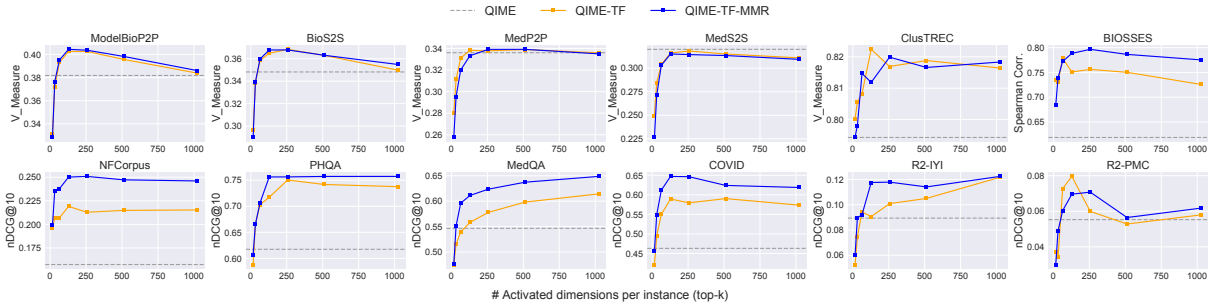

Figure 3: Effect of the top-k parameter in training-free embedding construction. We report performance on clustering (V-measure), STS (Spearman correlation), and retrieval (nDCG@10) benchmarks.

4.4 Effect of Top-k Dimension Activation in Training-Free Embedding Construction

Figure 3 examines how performance varies with the top-k selection parameter $k \in \{2^i \mid i = 3, \ldots, 10\}$ across clustering, STS, and retrieval tasks. We compare QIME-TF with QIME-TF-MMR, and QIME shown as a reference.

Across tasks, QIME-TF-MMR consistently matches or outperforms QIME-TF, with the largest gains on STS and retrieval benchmarks. Performance typically peaks at moderate values of k (around 128 or 256), after which improvements saturate or slightly decline due to redundancy among selected questions. Notably, QIME-TF-MMR often reaches or surpasses the performance of classifier-based QIME with only a few hundred active dimensions per instance, demonstrating that sparse, diversity-aware activation effectively balances efficiency, interpretability, and effectiveness.

4.5 Ablation Study

Table 4 examines the effect of medical ontology grounding in QIME by comparing the full model with a variant that removes ontology grounding during question generation while keeping all other components fixed. Removing ontology grounding consistently degrades performance for both classifier-based and training-free variants across similarity, clustering, and retrieval benchmarks. This confirms that ontology grounding is a critical component of QIME, enabling more informative and discriminative question dimensions.

5 Conclusion

We introduce QIME, an ontology-grounded framework for constructing question-based interpretable medical text embeddings. By grounding dimension generation in UMLS concept signatures, QIME produces clinically relevant and semantically discriminative representations. Experiments show that QIME consistently outperforms prior interpretable models and narrows the gap to black-box biomedical encoders across clustering, semantic similarity, and retrieval tasks. Its training-free construction enables efficient, sparse, and self-describing embeddings, offering an effective and practical foundation for transparent medical NLP systems.

Limitations

Despite its effectiveness, QIME has several limitations. First, the quality of the learned question dimensions depends on the coverage and accuracy of both the underlying medical corpus and the medical ontology; incomplete, outdated, or noisy concept inventories may limit performance or introduce spurious dimensions in rapidly evolving domains. Second, QIME produces interpretable embeddings grounded in general medical concepts, but interpretability requirements can differ across user groups, such as biomedical researchers, clinical practitioners, or policy analysts. Designing audience-specific interpretable representations and systematically evaluating their utility in real clinical workflows remain important directions for future work.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *CoRR*, abs/2404.05961.

Vinamra Benara, Chandan Singh, John X. Morris, Richard J. Antonello, Ion Stoica, Alexander Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking llms questions. In *NeurIPS*.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *ECIR*, volume 9626 of *Lecture Notes in Computer Science*, pages 716–722. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Kenneth C. Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, and 1 others. 2025. MMTEB: massive multilingual text embedding benchmark. *CoRR*, abs/2502.13595.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910. Association for Computational Linguistics.

Diego Garcia-Olano, Yasumasa Onoe, Ioana Baldini, Joydeep Ghosh, Byron C. Wallace, and Kush R. Varshney. 2021. Biomedical interpretable entity representations. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3547–3561. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinform.*, 39(10).

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. In *EMNLP (Findings)*, pages 8838–8855. Association for Computational Linguistics.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Lei Li, Xiao Zhou, and Zheng Liu. 2025. R2MED: A benchmark for reasoning-driven medical retrieval. *CoRR*, abs/2505.14558.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *NAACL-HLT*, pages 4228–4238. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *EACL*, pages 2006–2029. Association for Computational Linguistics.

Juri Opitz, Lucas Moeller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22303–22319. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *EMNLP*, pages 1532–1543.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *J. Am. Medical Informatics Assoc.*, 31(9):1844–1855.

Richard J. Roberts. 2001. Pubmed central: The genbank of the published literature. *Proceedings of the National Academy of Sciences*, 98(2):381–382.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony Kum Hoe Tung, and Jun Yu. 2025. A general framework for producing interpretable semantic text embeddings. In *ICLR*. OpenReview.net.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *CoRR*, abs/2509.20354.

Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1:1–1:12.

Yile Wang, Zhanyu Shen, and Hui Huang. 2025. LDIR: low-dimensional dense and interpretable text embeddings with relative representations. In *ACL (Findings)*, pages 14397–14409. Association for Computational Linguistics.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinform.*, 37(17):2792–2794.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. In *EMNLP*, pages 22234–22254. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

A Prompt Templates

We use large language models to generate contrastive, ontology-grounded yes/no questions for each semantic cluster. The prompt below illustrates the template used for contrastive question generation.

Prompt: Contrastive Question Generation

You are an expert in biomedical NLP question generation. Generate 10 simple yet insightful yes/no questions that determine properties of an article, where for all questions, the answer will be "yes" for ALL the positive articles and "no" for ALL the negative articles. Questions must focus on biomedical meaning, such as diseases, symptoms, risk factors, treatments, drugs, genes and chemicals.

Constraints:

- Avoid **meta-level** questions (e.g., "Does the article report results from a clinical trial?", "Does the article discuss methods?").
- IMPORTANT: Output ONLY the numbered questions, no analysis or explanation.
- Format the questions in a numbered list as shown below: 1. [First simple yes/no question] 2. [Second simple yes/no question]

Positive Articles:
Positive 1. {positive_chunk_1}
Positive 2. {positive_chunk_2}
...
Positive N. {positive_chunk_N}

Negative Articles:
Negative 1. {negative_chunk_1}
Negative 2. {negative_chunk_2}
...
Negative M. {negative_chunk_M}

UMLS Context:
CUI 1: description_1
CUI 2: description_2
...

B Post-Processing of Generated Questions

To ensure discriminative and reliable question dimensions, we apply a post-processing and filtering procedure to the generated questions.

Sampling Strategy. For each cluster, we sample $p_{\text{p}} = 5$ positive documents from the cluster, $p_{\text{hard}} = 3$ hard negatives from the nearest clusters, and $p_{\text{easy}} = 2$ easy negatives from random corpus positions.

Answer Probing. An LLM is used to answer each question for all sampled documents. Responses are normalized to binary `yes`/`no` labels.

Discrimination Scoring. Each question is assigned a discrimination score:

$$\text{score} = \frac{\text{yes}_{\text{pos}}}{\text{yes}_{\text{pos}} + \text{no}_{\text{pos}}} - \frac{\text{yes}_{\text{neg}}}{\text{yes}_{\text{neg}} + \text{no}_{\text{neg}}}. \quad (1)$$

Questions are ranked by this score in descending order.

Redundancy Filtering. To remove near-duplicate questions, we compute cosine similarity between question embeddings and retain only questions with similarity below a threshold $\theta = 0.8$. For each cluster, up to $adapt_{\text{t}}$ questions are selected, where $adapt_{\text{t}}$ scales with cluster size. Finally, questions are deduplicated across clusters to form the global question set used for embedding construction.

C Training Per-Question Classifiers

In the classifier-based approach, we associate each embedding dimension with a binary classifier corresponding to a question. For each question, we construct 1,000 training instances by sampling 300 positive examples from the corresponding cluster, 500 hard negatives from the nearest clusters, and 200 random negatives.

We attach one classification head per question on top of a shared backbone encoder, freeze the backbone parameters, and train only the classification heads. Training is formulated as a multi-task classification problem: given a document–question pair, only the corresponding head is updated using the cross-entropy loss. We train for 3 million steps with a batch size of 1. After training, the outputs of all classification heads for one input text are concatenated to form the final classifier-based embedding.