Asking the Right Questions: Ontology-Grounded Interpretable Embeddings for Medical Text

Anonymous ACL submission

Abstract

While dense biomedical embeddings achieve strong performance, their black-box nature limits transparency in high-stakes clinical settings. Recent question-based interpretable embeddings represent text through binary answers to natural-language questions, but existing approaches often rely on corpus-driven signals and may overlook clinically meaningful distinctions. We propose **QIME**, an ontology-grounded framework for interpretable medical text embeddings in which each dimension corresponds to a clinically meaningful yes/no question. QIME leverages biomedical ontologies to guide question discovery from semantic clusters, producing atomic questions grounded in domain knowledge. It further introduces a similarity-based semantic activation mechanism with diversity-aware dimension selection to construct sparse, interpretable representations. Experiments on 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, providing concise and clinically informative explanations.

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, which limits transparency in error analysis and clinical auditing.

To address this limitation, 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, increasing 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 provides explicit semantic dimensions, generating clinically meaningful questions from unstructured corpora remains challenging, often capturing topical or stylistic differences rather than fine-grained medical distinctions. Moreover, existing approaches typically incur substantial computational overhead due to extensive LLM queries or large numbers of supervised classifiers.

Therefore, we build on three key 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 critically affect the interpretability of the resulting embeddings. Third, practical deployment requires embedding construction mechanisms that remain computationally efficient without relying on large-scale supervision or expensive LLM-based inference. Figure 1 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 discovery process. Specifically, we cluster a large medical corpus and ex-

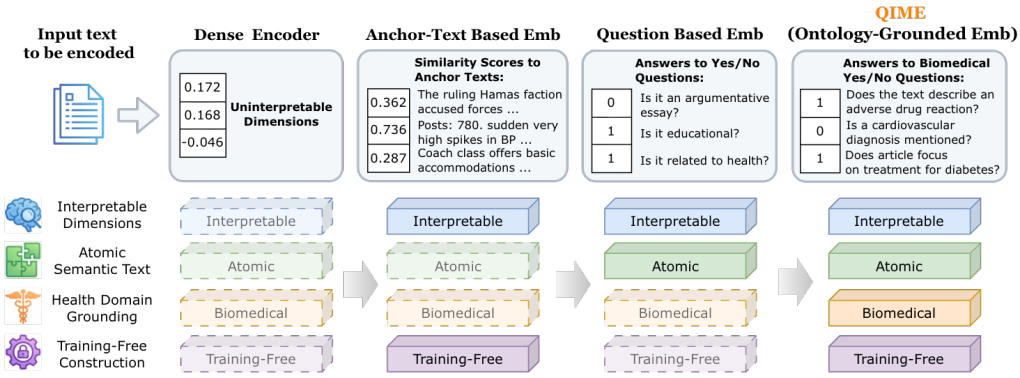

Figure 1: Comparison of text embedding paradigms. Dense encoders produce uninterpretable vectors, anchor-based embeddings rely on heterogeneous reference texts, and question-based embeddings use binary questions derived from corpus-driven signals. **QIME** grounds question dimensions in biomedical ontology concepts and constructs interpretable embeddings via similarity-based semantic activation with diversity-aware dimension selection.

tract biomedical concept signatures for each cluster, which are used to guide an LLM to generate discriminative, domain-specific questions. In the second stage, QIME constructs sparse interpretable embeddings using a similarity-based semantic activation mechanism that measures the semantic alignment between input texts and question dimensions. To reduce redundancy among activated questions, we further incorporate diversity-aware dimension selection via Maximal Marginal Relevance (MMR). This design eliminates the need to train large numbers of per-question classifiers and enables scalable construction of interpretable embeddings.

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 with clinically meaningful and discriminative dimensions.
- We introduce a **similarity-based semantic activation** mechanism for constructing sparse interpretable embeddings, together with diversity-aware dimension selection to reduce redundancy among activated questions.
- We demonstrate that QIME achieves **strong empirical performance and interpretability** across biomedical similarity, clustering, and retrieval benchmarks.

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.

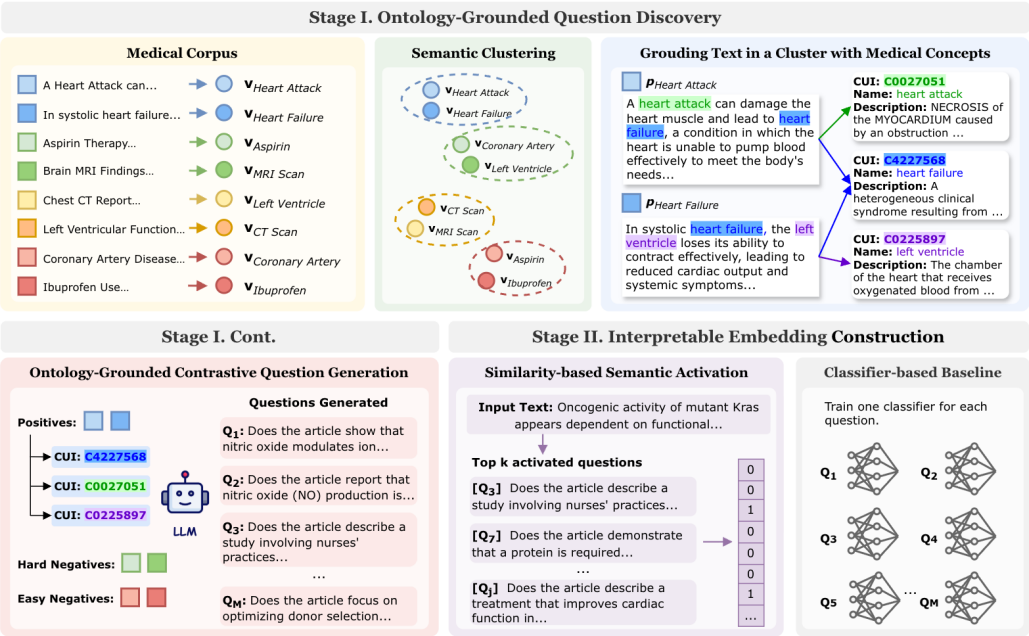

Figure 2: Overview of the QIME framework.

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. However, embedding construction relies on training a dedicated classifier for each dimension, introducing 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 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, enabling the discovery of semantically atomic and clinically meaningful question dimensions. QIME then constructs sparse interpretable embeddings through a similarity-based semantic activation mechanism with diversity-aware selection.

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, they are automatically discovered through an ontology-grounded process. A contrastive prompting strategy is used to generate discriminative candidate questions.

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 Discovery

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 first cluster the corpus into semantically coherent groups. Each document x_i is encoded into a dense representation $\mathbf{h}_i$. The corpus is then clustered into K clusters, $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{C}_k$. Each cluster typically corresponds to 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 specific 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 $\mathcal{U}_k = \{u_1, u_2, \ldots, u_{|\mathcal{U}_k|}\}$.

Ontology-Grounded Question Discovery 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 generate candidate questions using a contrastive prompting strategy, while grounding the generation process in biomedical ontology concepts extracted from each cluster to ensure medical relevance.

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

1. **Positive samples** Documents drawn from $\mathcal{C}_k$.
2. **Hard negatives** Documents from clusters that are semantically proximate to $\mathcal{C}_k$.
3. **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 through a similarity-based semantic activation mechanism.

Similarity-Based Semantic Activation. Unlike prior approaches that require answering every question, QIME activates only the top-k relevant dimensions. QIME constructs embeddings through a similarity-based semantic activation mechanism that measures the semantic alignment between an input document and each question dimension. Each question acts as an interpretable semantic probe, and dimensions are activated based on their relevance to the document, producing sparse and interpretable representations without requiring explicit QA supervision.

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

We then compute cosine similarity scores $s_j = \mathrm{sim}(\mathbf{h}(x), \mathbf{h}(q_j))$, which measure the semantic relevance between the document and each question. The embedding activates the most relevant question dimensions through a top-k selection, where

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
Inter-pretable	**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	**40.37**	**36.78**	**33.92**	**31.44**	**81.99**	**44.90**	**79.66**

Table 1: Clustering (V-measure) and semantic textual similarity (Spearman correlation) results on biomedical benchmarks.

typically $k \ll M$, resulting in sparse interpretable representations:

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

To reduce redundancy among highly similar questions, QIME further incorporates maximal marginal relevance (MMR) during the selection process. Instead of selecting dimensions solely based on similarity to the document, questions are chosen iteratively by jointly maximizing relevance to the document and dissimilarity to previously selected questions. This encourages the activated dimensions to capture complementary semantic aspects of the input.

By activating a small set of diverse questions, QIME produces concise and interpretable embeddings while avoiding the need to train large numbers of per-question classifiers or collect question–answer supervision.

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 MedicalQARetrieval (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 Embed-

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
Inter-pretable	**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	25.09	**75.64**	62.36	**64.65**	11.79	7.08	**41.10**

Table 2: Retrieval results (nDCG@10) across biomedical benchmarks.

dingGemma (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 the MMR-based dimension selection is set to 0.7. The value of k in the top-k dimension activation is set to 256.

4.2 Main Results

Clustering and STS. Table 1 reports results on clustering and STS benchmarks. Among black-box encoders, domain-specialized models such as MedEmbed, EmbeddingGemma, and BioLORD achieve the strongest performance, reflecting the benefits of large-scale biomedical pretraining.

Within the interpretable category, QIME consistently outperforms interpretable baselines across all clustering tasks and on BIOSSES. In particular, QIME substantially improves over QA-Emb and CQG-MBQA, indicating that ontology-grounded question discovery produces more coherent and discriminative semantic dimensions than manually designed or purely data-driven questions.

Retrieval. Table 2 reports retrieval performance measured by nDCG@10 on biomedical benchmarks. While supervised dense encoders such as MedEmbed achieve the strongest overall performance, QIME attains the best results among interpretable methods and remains competitive on challenging datasets including PHQA, MedQA, and TREC–COVID.

Additional experiments with alternative LLM backbones (LLaMA-3-70B and GPT-5-mini) show similar performance trends, indicating that QIME is robust to the choice of backbone (Appendix D).

Overall, QIME substantially narrows the gap between interpretable and dense embeddings, achieving a favorable balance between effectiveness and interpretability. In addition, the similarity-based activation mechanism further avoids training large numbers of per-question classifiers, making the

Model	Clustering (V-Measure ↑)						STS (SC ↑)
	BioP2P	BioS2S	MedP2P	MedS2S	Covid	Average	BIOSSES
QIME	40.36	36.78	33.92	31.44	81.99	44.90	79.66
QIME w/o MMR	40.26	36.83	33.78	31.83	81.69	44.88	75.60
QIME w/o ontology	38.75	35.60	33.36	31.21	81.21	44.03	74.32
QIME_CLS	38.18	34.82	33.61	32.00	79.43	43.61	61.88
QIME_CLS w/o ontology	37.29	34.01	32.91	31.12	79.22	42.91	57.77
	Retrieval (nDCG@10 ↑)						
	NFCorpus	PHQA	MedQA	COVID	R2-IYI	R2-PMC	Average
QIME	25.09	75.64	62.36	64.65	11.79	7.08	41.10
QIME w/o MMR	21.29	75.04	57.79	57.96	10.09	5.99	38.03
QIME w/o ontology	20.76	74.13	58.70	54.29	11.00	7.27	37.69
QIME_CLS	15.74	61.74	54.66	46.31	8.94	5.53	32.15
QIME_CLS w/o ontology	15.35	61.07	51.45	44.56	8.82	4.39	30.94

Table 3: Ablation study analyzing the impact of ontology grounding, MMR-based diversity selection, and classifier-based vs. similarity-based embedding construction.

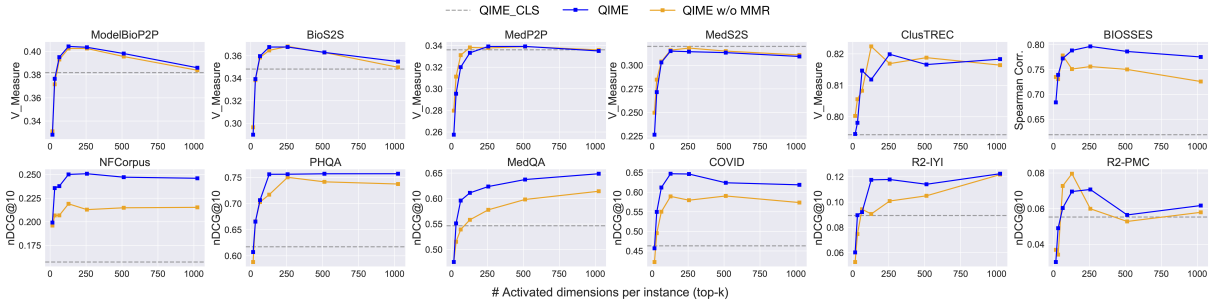

Figure 3: Effect of the top-k dimension activation in similarity-based semantic activation. Performance peaks at moderate sparsity (k128–256), and MMR-based selection consistently improves results.

framework efficient and practical for deployment.

4.3 Ablation Study

Table 3 analyzes three components of QIME: the embedding construction strategy, ontology grounding during question generation, and diversity-aware dimension selection via MMR. To isolate these factors, we evaluate five variants. In addition to the full model (QIME), we consider:

(i) QIME (w/o MMR), which removes diversity-aware selection;

(ii) QIME (w/o ontology), which disables ontology grounding during question generation;

(iii) QIME_CLS, a classifier-based embedding construction strategy where each dimension is predicted by a dedicated binary classifier, Implementation details are provided in Appendix C; and

(iv) QIME_CLS (w/o ontology), which further removes ontology grounding in QIME_CLS.

Effect of Embedding Construction Strategy. We first compare the proposed similarity-based semantic activation with the classifier-based variant QIME_CLS. Across clustering, STS, and retrieval tasks, QIME consistently outperforms QIME_CLS by a large margin. This suggests that directly measuring semantic alignment between documents and question dimensions is more effective than training independent per-question classifiers.

Effect of Ontology Grounding. Removing ontology grounding consistently degrades performance for both QIME and QIME_CLS. This confirms that grounding question generation in medical ontologies produces more informative and clinically meaningful question dimensions. Without ontology constraints, generated questions tend to capture more generic corpus-level patterns rather than domain-relevant distinctions.

Effect of Diversity-Aware Dimension Selection. Comparing QIME with QIME (w/o MMR) shows that diversity-aware selection yields small but consistent improvements across tasks, indicating that MMR helps reduce redundancy among activated questions and encourages the embedding to capture complementary semantic aspects of the input.

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	Rank	Content of the top-rated Dimensions
LDIR-500	1	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.
	2	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.
	3	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	1	Does the article discuss medical conditions or treatments?
	2	Is the article devoid of promoting extreme diets or quick fixes?
	3	Does the article provide factual information rather than anecdotal stories?
QIME	1	Is there a focus on pain control in terminally ill or cancer patients?
	2	Does the article involve the use of computed tomography (CT) for diagnosing cardiovascular conditions in humans?
	3	Is there evidence of a bacterial infection affecting heart valves or cardiac tissue?

Table 4: Qualitative comparison of the top-ranked embedding dimensions produced by different interpretable methods for the same clinical input. LDIR returns long anchor texts with irrelevant content and limited interpretability, while CQG-MBQA produces generic questions. In contrast, QIME activates concise, clinically grounded questions that directly reflect key aspects of the input.

4.4 Effect of Top-k Dimension Activation

Figure 3 examines how performance varies with the top-k activation parameter $k \in \{2^i \mid i = 3, \ldots, 10\}$ across clustering, STS, and retrieval tasks. We compare the full QIME model with a variant without MMR (QIME w/o MMR), while the classifier-based variant QIME_CLS is shown for reference.

Across tasks, QIME consistently matches or outperforms the version without MMR, indicating that diversity-aware selection improves dimension coverage. Performance remains stable across a broad range of k, typically peaking around 128–256. To ensure fair comparison and avoid dataset-specific tuning (since many embedding benchmarks lack development sets), all reported results use a fixed $k = 256$. Notably, QIME often matches or surpasses the classifier-based variant QIME_CLS using only a few hundred active dimensions, showing that sparse, diversity-aware activation provides an effective balance between performance and interpretability.

4.5 Case Study: Interpreting Question-Based Representations

Table 4 compares the top-ranked embedding dimensions produced by different interpretable methods for the same clinical input describing chest pain in a lung cancer patient.

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 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.

5 Conclusion

We introduce **QIME**, an ontology-grounded framework for constructing question-based interpretable medical text embeddings. By grounding question generation in UMLS concept signatures, QIME discovers clinically meaningful semantic dimensions. To construct embeddings from these questions, QIME employs a similarity-based semantic activation mechanism with diversity-aware dimension selection, enabling sparse and semantically coherent representations. Experiments across clustering, semantic similarity, and retrieval benchmarks demonstrate that QIME consistently outperforms prior interpretable embedding methods and substantially narrows the performance gap to strong black-box biomedical encoders. These results suggest that ontology-guided question discovery combined with semantic activation provides an effective foundation for interpretable and transparent medical NLP systems.

Limitations

QIME relies on external biomedical resources: a corpus for clustering and an ontology for grounding question generation. As with other knowledge-grounded NLP systems, the coverage and currency of these resources may influence the resulting question dimensions. In our experiments, we use PubMed and UMLS, which provide broad biomedical coverage, though these resources continue to evolve over time. Importantly, QIME is modular: updating the ontology or corpus only requires regenerating the question set, without retraining dense encoders.

Second, there remains a gap between QIME and dense encoders on some retrieval tasks. This reflects a design trade-off: while sparse question-based embeddings may capture less fine-grained signals than dense representations, they provide transparent, clinically meaningful dimensions that support interpretability and auditability in biomedical applications.

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 ontology-grounded question discovery from semantic clusters

Prompt: Ontology-Grounded Question Discovery

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 $p_{\mathrm{p}} = 5$ positive documents from the cluster, $p_{\mathrm{hard}} = 3$ hard negatives from the nearest clusters, and $p_{\mathrm{easy}} = 2$ 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 $adapt_{\mathrm{t}}$ questions are selected, where $adapt_{\mathrm{t}}$ scales with cluster size. Finally, questions are deduplicated across clusters to form the global question set used for embedding construction.

C Classifier-Based Embedding Construction (QIME_CLS)

As a comparison to the proposed similarity-based semantic activation, we implement a classifier-based embedding construction strategy, denoted as **QIME_CLS**. In this formulation, each question dimension is modeled as a binary prediction task.

Given a document x and a question q_j, the embedding value for dimension j is predicted by a dedicated binary classifier that determines whether the semantic property expressed by q_j is present in x. This approach replaces semantic activation of **QIME** with supervised classification.

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,

Model	Backbone	Clustering (V-Measure ↑)						STS (SC ↑)
		BioP2P	BioS2S	MedP2P	MedS2S	ClusTREC	Average	BIOSSES
QIME_CLS	Qwen3-30B	38.18	34.82	33.61	32.00	79.43	43.61	61.88
QIME	Qwen3-30B	40.37	36.78	33.92	31.44	81.99	44.90	79.66
QIME_CLS	LLaMA-3-70B	37.65	34.29	33.18	31.37	80.53	43.41	60.55
QIME	LLaMA-3-70B	40.24	35.92	33.90	31.59	80.86	44.50	78.48
QIME_CLS	GPT-5-mini	36.75	34.07	33.41	31.39	80.69	43.26	59.65
QIME	GPT-5-mini	40.09	36.67	34.08	31.79	80.96	44.72	75.05

Table 5: Clustering and STS performance across different LLM backbones used for question generation. Performance remains consistent across backbones.

Model	Backbone	Retrieval (nDCG@10 ↑)						
		NFCorpus	PHQA	MedQA	COVID	R2-IYI	R2-PMC	Average
QIME_CLS	Qwen3-30B	15.74	61.74	54.66	46.31	8.94	5.53	32.15
QIME	Qwen3-30B	25.09	75.64	62.36	64.65	11.79	7.08	41.10
QIME_CLS	LLaMA-3-70B	16.15	64.95	55.51	47.57	8.78	5.14	33.02
QIME	LLaMA-3-70B	24.85	75.30	62.91	63.47	11.54	8.12	41.03
QIME_CLS	GPT-5-mini	16.48	66.20	57.04	47.73	9.27	5.27	33.66
QIME	GPT-5-mini	25.32	73.42	61.47	60.92	11.16	6.73	39.84

Table 6: Retrieval performance across LLM backbones used for question generation. Results remain stable across backbones.

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 a given input text are concatenated to form the final classifier-based embedding.

D Backbone Sensitivity Analysis

To evaluate the sensitivity of QIME to the choice of LLM backbone used for question generation, we additionally generate question sets using LLaMA-3-70B and GPT-5-mini, while the main paper reports results using Qwen3-30B.

Tables 5 and 6 report the results across clustering, semantic textual similarity (STS), and retrieval benchmarks. Across all tasks, performance remains highly consistent across different LLM backbones. In particular, QIME maintains similar improvements over the classifier-based variant regardless of the backbone used.

These results suggest that the performance gains of QIME primarily arise from ontology-grounded question construction and similarity-based semantic activation with diversity-aware dimension selection, rather than reliance on a specific LLM backbone.