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 rely on corpus-driven signals rather than clinically grounded criteria, often capturing topical or stylistic differences rather than fine-grained medical 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 contrastive question generation from semantic clusters, producing atomic, domain-grounded questions. It constructs embeddings via similarity-based semantic activation with MMR-based diversity-aware dimension selection, yielding sparse representations efficiently. 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. It also imposes significantly lower cognitive burden than existing methods, providing concise and clinically informative interpretations.

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.

Consider a concrete clinical scenario: a retrieval system that surfaces similar historical patient cases to support decision-making. With dense embeddings, the system returns only a scalar score with no insight into *why* a case is retrieved. With interpretable embeddings, clinicians can directly inspect the semantic dimensions driving the result. For example, *"Does this case involve CT-based cardiovascular diagnosis?"* or *"Is there evidence of metastatic disease?"* reveal the semantic basis of the retrieval. Such transparency is critical for clinical auditing and system refinement: when an expected case is not retrieved or an irrelevant case surfaces, interpretable dimensions allow clinicians to identify which semantic properties are misaligned and correct them accordingly to improve the retrieval system, which is fundamentally impossible with dense representations.

To address this limitation, a growing body of work has explored interpretable text embeddings with human-understandable dimensions. Early efforts include Concept Bottleneck Models (CBMs) (Koh et al., 2020), which use predefined concepts as intermediate representations. Anchor-based methods (Wang et al., 2025) represent texts by similarity to reference documents, but require users to inspect heterogeneous anchors. Recently, question-based embeddings (Sun et al., 2025; Benara et al., 2024) define each dimension as a binary natural-language question. While this paradigm provides explicit semantic dimensions, generating clinically meaningful questions from unstructured corpora remains difficult, often yielding topical or stylistic dimensions rather than fine-grained medical distinctions. Existing methods also incur substantial computational overhead due to extensive LLM calls or massive supervised classifiers.

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 struc-

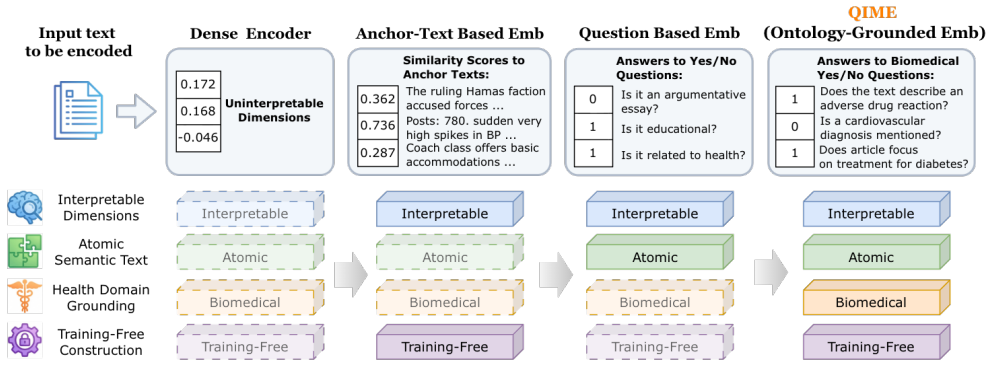

Figure 1: Comparison of text embedding paradigms. Dense encoders produce opaque vectors, anchor-based embeddings rely on heterogeneous reference texts, and question-based embeddings use binary questions derived from corpus-driven signals. **QIME** grounds question dimensions in biomedical ontology concepts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

tured medical knowledge and interpretable natural-language representations through ontology-grounded question discovery. Specifically, we cluster a large medical corpus, extract biomedical concept signatures for each cluster, and guide an LLM to generate discriminative, domain-specific questions. Next, QIME constructs sparse interpretable embeddings by activating semantically aligned questions as representation coordinates. To reduce redundancy, we further incorporate diversity-aware dimension selection via Maximal Marginal Relevance (MMR). This avoids training massive per-question classifiers and enables scalable construction of interpretable embeddings. Figure 1 shows the distinctions among embedding paradigms.

We evaluate QIME on biomedical benchmarks spanning semantic textual similarity, clustering, and information retrieval. Results show that QIME outperforms prior interpretable embedding methods and substantially narrows the gap to strong black-box biomedical encoders. Qualitative analyses further demonstrate that QIME provides explicit and clinically grounded dimensions.

Our main contributions are:

- We propose **QIME**, an **ontology-grounded** framework for constructing **question-based interpretable medical text embeddings** with clinically meaningful dimensions.
- We introduce a **similarity-based semantic activation** mechanism for constructing sparse interpretable embeddings, with **diversity-aware dimension selection** to reduce redundancy among activated questions.
- We show that QIME achieves **strong empirical performance** across biomedical similarity, clustering, and retrieval benchmarks, and **reduces cognitive load** by over 76% compared to existing question-based encoders.

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., 2021). Ontology-aware models, SapBERT (Liu et al., 2021) and BioLORD (Remy et al., 2024), leverage the Unified Medical Language System (UMLS) (Bodenreider, 2004) 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 when new concepts are added or modified.

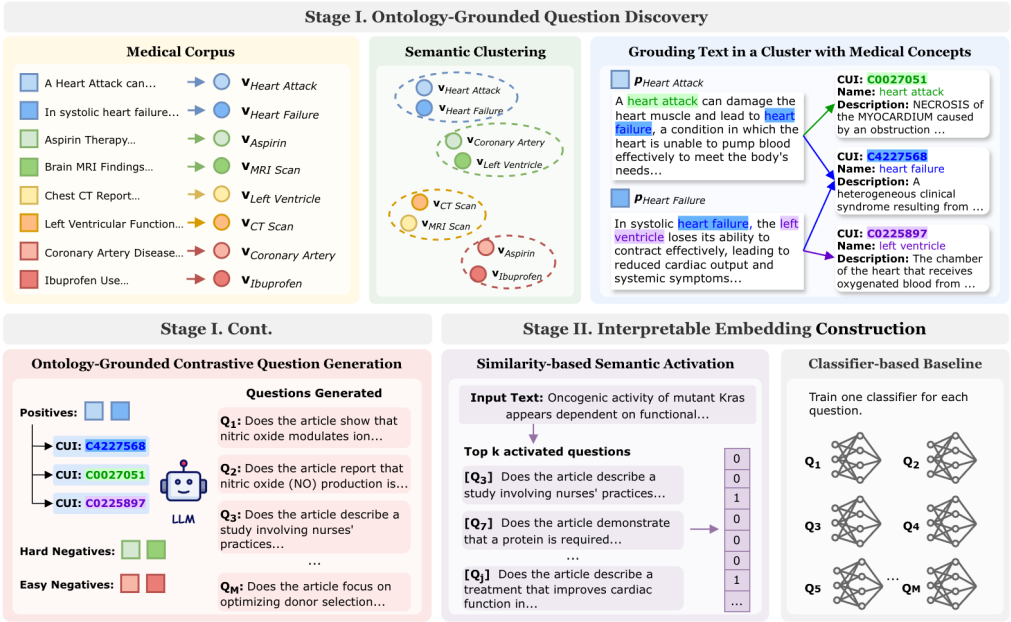

Figure 2: Overview of the QIME framework.

A more recent direction is question-based embeddings, where dimensions are associated with binary natural-language questions. 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, but relies solely on corpus-level distributional signals, yielding generic questions that lack clinical specificity. Moreover, it 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 anchors, achieving compact representations but requiring users to interpret dimensions through long, heterogeneous anchor texts rather than self-describing semantic units.

In contrast, QIME grounds question generation in UMLS concept signatures, constructs embeddings via training-free semantic activation, and applies diversity-aware dimension selection via MMR to yield sparse, complementary, and clinically interpretable representations.

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 yes/no question (e.g., *"Does the text describe adverse drug reactions?"*). Formally, given a medical text corpus $\mathcal{D} = \{x_i\}_{i=1}^{N}$, QIME aims to learn an embedding function $f : x \mapsto \mathbf{z} \in \{0,1\}^M$, where $z_j \in \{0,1\}$. The binary value indicates whether a question dimension is semantically activated for the document. Unlike prior question-based approaches, questions are not predefined or heuristic but automatically discovered through an ontology-grounded contrastive prompting process.

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; (2) **Interpretable embedding construction**, which encodes new texts into sparse question-indexed representations via similarity-based semantic activation and diversity-aware dimension selection through MMR, without requiring QA supervision or classifier training.

3.2 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.

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), a large-scale, publicly available biomedical ontology, 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 binary medical questions capturing the cluster's defining semantic properties via a contrastive prompting strategy. For each cluster, we construct three types of examples: **positive samples** drawn from $\mathcal{C}_k$; **hard negatives** from semantically nearest clusters (by centroid distance); and **easy negatives** from randomly selected distant clusters, sampling six examples from each selected cluster.

An LLM is prompted to generate yes/no questions that distinguish positives from both negative types, 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. Generated questions are aggregated and post-processed to remove low-quality, ambiguous, and redundant entries, resulting in a final set of M questions, $\mathcal{Q} = \{q_1, \ldots, q_M\}$. Prompts and post-processing details are in Appendix A and B.

3.3 Interpretable Medical Embedding Construction

Once the question set $\mathcal{Q}$ is obtained, the second stage constructs interpretable embeddings for individual documents.

Similarity-Based Semantic Activation. 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 encode all questions into $\{\mathbf{h}(q_j)\}_{j=1}^{M}$. We compute cosine similarity scores $s_j = \mathrm{sim}(\mathbf{h}(x), \mathbf{h}(q_j))$ and activate the top-k dimensions, where $k \ll M$:

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

To reduce redundancy among highly similar questions, QIME further incorporates Maximal Marginal Relevance (MMR) during the selection process: $q^* = \arg\max_{q_i \in Q \setminus S} \Big[\lambda \cdot \mathrm{sim}(q_i, d) - (1-\lambda) \cdot \max_{q_j \in S} \mathrm{sim}(q_i, q_j)\Big]$, where S is the set of already-selected questions. Questions are chosen iteratively by jointly maximizing relevance to the document and dissimilarity to previously selected questions, encouraging activated dimensions to capture complementary semantic aspects of the input. By activating a small set of diverse questions, QIME produces concise and interpretable embeddings.

Computational Efficiency. Compared to prior question-based methods, QIME's embedding construction is substantially more efficient. QA-Emb requires M separate LLM queries at inference time, while classifier-based methods such as CQG-MBQA require training M dedicated classifiers (requiring millions of annotated instances) and querying all M at inference. QIME requires no training, and inference consists of a single encoder forward pass, one matrix multiplication between the document and all M question embeddings, and $O(k \times M)$ lightweight cosine computations for MMR selection.

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. Best results are in bold and second-best results are underlined within each category.

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) text retrieval, following the widely adopted text embedding evaluation protocol established by MTEB (Muennighoff et al., 2023). The tasks jointly assess the full spectrum of embedding quality: 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). They require grouping biomedical preprints based on either titles or abstracts. 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 (Zhang et al., 2026), 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., 2021), 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 (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.

Implementation Details. We preprocess the widely adopted PubMed corpus (Roberts, 2001), randomly sampling 5M out of 25M paragraphs (average length 296 tokens) for semantic cluster-

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.

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 each component of QIME.

ing after filtering low-quality and duplicated entries. Embeddings are obtained using MedEmbed (Balachandran, 2024) and clustered by k-means (k=2,500). Medical entities are extracted using HunFlair (Weber et al., 2021). For the cluster-level concept signature, we retain the top-5 most frequent CUIs for each cluster as $\mathcal{U}_k$ to ensure focus. Qwen3-30B (Yang et al., 2025) is used as the LLM backbone for grounded question generation. After post-processing, 8,855 questions are retained. We set $\lambda = 0.7$ in MMR and $k = 256$ in the top-k dimension activation.

4.2 Main Results: Embedding Quality

Clustering and STS. Table 1 reports results on clustering and STS benchmarks. Domain-specialized black-box models such as MedEmbed, EmbeddingGemma, and BioLORD achieve the strongest performance. Notably, QIME surpasses all black-box encoders on 4 out of 5 clustering tasks, a strong outcome for an interpretable method.

Within the interpretable category, QIME consistently outperforms interpretable baselines across all clustering and STS tasks. 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 (nDCG@10). While supervised dense encoders such as MedEmbed achieve the strongest overall performance, QIME attains the best average results among all interpretable methods, outperforming CQG-MBQA on every dataset. It falls marginally

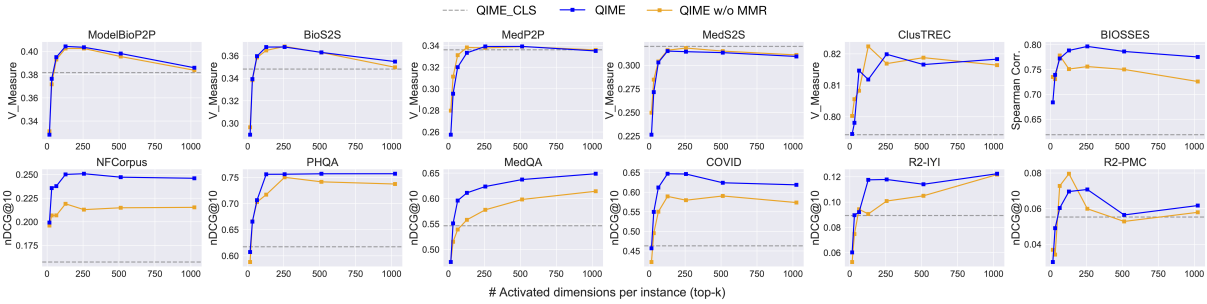

Figure 3: Effect of the top-k dimension selection in similarity-based semantic activation. Performance peaks at moderate sparsity (k=128–256), and MMR-based selection consistently improves results.

behind LDIR-500 on some datasets, though LDIR-500 offers substantially less interpretability via heterogeneous anchor texts.

Overall. Across all three task types, QIME outperforms prior interpretable methods and substantially narrows the gap to dense embeddings, without per-question classifier training or test-time LLM inference. Additional experiments with alternative LLM backbones (LLaMA-3-70B and GPT-5-mini) show similar trends, indicating that QIME is robust to the choice of backbone (Appendix D).

4.3 Ablation Study

Table 3 analyzes three components of QIME: embedding construction strategy, ontology grounding for question generation, and diversity-aware dimension selection via MMR. To isolate these factors, we evaluate four variants: (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 (details 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 achieves better average results than QIME_CLS, indicating the effectiveness of directly measuring semantic alignment between documents and question dimensions.

Effect of Ontology Grounding. Removing ontology grounding causes performance drops on the majority of tasks for both QIME and QIME_CLS. This confirms that grounding question generation in medical ontologies produces more informative and clinically meaningful question dimensions.

Effect of Diversity-Aware Dimension Selection. Comparing QIME with QIME w/o MMR shows that diversity-aware selection improves performance on 10 out of 12 tasks, indicating that MMR reduces redundancy among activated questions and encourages the embedding to capture complementary semantic aspects of the input.

4.4 Effect of Top-k Dimension Activation

Figure 3 examines how performance varies with the top-k activation parameter $k \in \{2^i \mid i = 3, \ldots, 10\}$. We compare the full QIME model with QIME w/o MMR, while the classifier-based variant QIME_CLS is shown for reference.

Across tasks, QIME's performance remains stable across a broad range of k, typically peaking around 128–256. All reported results use a fixed $k = 256$ to ensure fair comparison across benchmarks. Notably, QIME often matches or surpasses the classifier-based variant using only a few hundred active dimensions, showing that sparse, diversity-aware activation provides an effective balance between performance and interpretability.

5 Interpretability Analysis

Beyond embedding quality, we evaluate the interpretability of QIME's representations through cognitive load analysis and a qualitative case study.

5.1 Cognitive Load

Following CQG-MBQA (Sun et al., 2025), we measure cognitive load as the expected number of shared activated dimensions between two documents, corresponding to the inner product of their binary embedding vectors. A lower cognitive load

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 top-ranked interpretable dimensions across methods for the same clinical input. LDIR returns lengthy, noisy anchor texts, while CQG-MBQA produces generic questions. QIME activates concise, clinically grounded questions that are easier to inspect and more closely aligned with clinical themes of the input.

Model	NFCorpus	PHQA	MedQA	COVID	R2-IYI	R2-PMC	Average
QA-Emb	1,515	1,883	1,682	2,071	2,191	2,106	1,908
CQG-MBQA	479	512	529	543	722	713	583
QIME	101	115	121	150	180	152	136

Table 5: Cognitive load (lower is better) across retrieval datasets. QIME imposes substantially less cognitive burden than existing question-based interpretable encoders.

indicates fewer dimensions are required to explain a similarity judgment.

Table 5 reports cognitive load across all retrieval datasets for question-based interpretable methods. QIME achieves substantially lower cognitive load than all baselines, averaging 136 shared dimensions per document pair compared to 583 for CQG-MBQA and 1,908 for QA-Emb, representing reductions of 76.7% and 92.9% respectively. This indicates that QIME's sparse representations impose less cognitive burden on users seeking to understand similarity judgments.

5.2 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 generic and fail to capture clinically specific distinctions. In contrast, QIME activates atomic, medically grounded questions that are easier to inspect and more closely aligned with the clinical themes of the input, such as CT-based cardiovascular diagnosis and cancer-related pathology.

6 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. Embeddings are then constructed via similarity-based semantic activation with diversity-aware dimension selection through MMR, yielding sparse and coherent representations without classifier training. Experiments across clustering, semantic similarity, and retrieval benchmarks demonstrate that QIME outperforms prior interpretable embedding methods and substantially narrows the performance gap to strong black-box biomedical encoders. Interpretability analysis further shows that QIME reduces cognitive load by over 76% relative to existing question-based encoders, providing an effective foundation for transparent and auditable clinical NLP applications.

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, activated questions in QIME indicate semantic relevance, which is the primary objective of embedding models, rather than factual entailment. They are intended to support embedding inspection and error analysis, while downstream clinical decision support would require additional task-specific tuning.

Finally, 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.

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. https://github.com/abhinand5/MedEmbed.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

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 Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. MMTEB: massive multilingual text embedding benchmark. In *ICLR*. OpenReview.net.

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. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*.

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.

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 2014–2037. Association for Computational Linguistics.

Juri Opitz, Lucas Möller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. In *EMNLP*, 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.

Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. 2026. R2MED: A benchmark for reasoning-driven medical retrieval. *CoRR*, abs/2505.14558.

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 the following post-processing procedure to the generated questions.

Sampling Strategy. For post-processing, we use a separate answer-probing sample to estimate whether each generated question discriminates the target cluster from negative examples. 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 randomly selected distant clusters.

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 cosine similarity below a threshold $\theta = 0.8$. For each cluster, up to $adapt_{\mathrm{t}}$ questions are retained, where $adapt_{\mathrm{t}}$ scales with cluster size: 2, 4, 6, and 8 questions for clusters of size <1k, 1k–2k, 2k–3k, and >3k, respectively, since larger clusters tend to encompass a broader range of medical semantics.

Global Deduplication. After per-cluster filtering, a final deduplication pass is applied across all clusters: any question with cosine similarity above θ=0.8 to any previously retained question is removed. This yields the final global question set $Q = \{q_1, \ldots, q_M\}$ 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.

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

Table 6: 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	Qwen3-30B	25.09	75.64	62.36	64.65	11.79	7.08	41.10
QIME_CLS	Qwen3-30B	15.74	61.74	54.66	46.31	8.94	5.53	32.15
QIME	LLaMA-3-70B	24.85	75.30	62.91	63.47	11.54	8.12	41.03
QIME_CLS	LLaMA-3-70B	16.15	64.95	55.51	47.57	8.78	5.14	33.02
QIME	GPT-5-mini	25.32	73.42	61.47	60.92	11.16	6.73	39.84
QIME_CLS	GPT-5-mini	16.48	66.20	57.04	47.73	9.27	5.27	33.66

Table 7: Retrieval performance across LLM backbones used for question generation. Results remain stable across backbones.

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, 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 6 and 7 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.