IMRNNs: An Efficient Method for Interpretable Dense Retrieval via Embedding Modulation

Anonymous ACL submission

Abstract

Interpretability in black-box dense retrievers remains a central challenge in Retrieval-Augmented Generation (RAG). Understanding how queries and documents semantically interact is critical for diagnosing retrieval behavior and improving model design. However, existing dense retrievers rely on static embeddings for both queries and documents, which obscures this bidirectional relationship. Post-hoc approaches such as re-rankers are computationally expensive, add inference latency, and still fail to reveal the underlying semantic alignment. To address these limitations, we propose Interpretable Modular Retrieval Neural Networks (`IMRNNs`), a lightweight framework that augments any dense retriever with dynamic, bidirectional modulation at inference time. `IMRNNs` employs two independent adapters: one conditions document embeddings on the current query, while the other refines the query embedding using corpus-level feedback from initially retrieved documents. This iterative modulation process enables the model to adapt representations dynamically and expose interpretable semantic dependencies between queries and documents. Empirically, `IMRNNs` not only enhance interpretability but also improve retrieval effectiveness. Across seven benchmark datasets, applying our method to standard dense retrievers yields average gains of +6.35% nDCG, +7.14% recall, and +7.04% MRR over state-of-the-art baselines. These results demonstrate that incorporating interpretability-driven modulation can both explain and enhance retrieval in RAG systems.

1 Introduction

Retrieval-Augmented Generation (RAG) systems have emerged as a dominant paradigm for grounding large language models (LLMs) in factual, domain-specific knowledge (Lewis et al., 2020; Glass et al., 2022). At the heart of these systems lies the *initial retriever*, responsible for selecting candidate documents from massive corpora before downstream re-ranking and generation. This component defines both the *efficiency* and *trustworthiness* of the entire pipeline. However, current dense retrieval methods face a fundamental limitation: they operate with *static embeddings* that encode queries and documents into fixed vector representations, preventing semantic adaptation between them at inference time.

This static nature creates two intertwined problems. First, it limits *retrieval performance*. When query and document embeddings cannot adapt to each other's semantic context, the retriever struggles to capture context-sensitive relevance signals, particularly for complex or ambiguous queries (Li et al., 2021). Second, and more critically underexplored, it lacks *interpretability*, unlike lexical methods (e.g., BM25) that provide transparent term-matching explanations (Robertson and Zaragoza, 2009a), dense retrievers function as black boxes, preventing users from understanding which aspects of a query influenced retrieval decisions or how specific document features contributed to ranking outcomes (Zhou et al., 2024). This opacity is particularly problematic in high-stakes domains such as healthcare (Munnangi et al., 2025), legal research (Magesh et al., 2025), and finance (Kim et al., 2025), where understanding retrieval decisions is essential for trust and accountability.

Recent research has begun addressing retrieval performance through lightweight adaptation mechanisms. *Retrieval adapters*, such as SEARCH-ADAPTOR (Yoon et al., 2024a) and HYPEN-CODER (Killingback et al., 2025a) have applied learned transformations over frozen embeddings, while DIME (Campagnano et al., 2025a) leverages Matryoshka representations to select important embedding subspaces. However, these methods share a critical gap: *they prioritize computational efficiency and marginal performance gains while providing no semantic-level interpretability*.

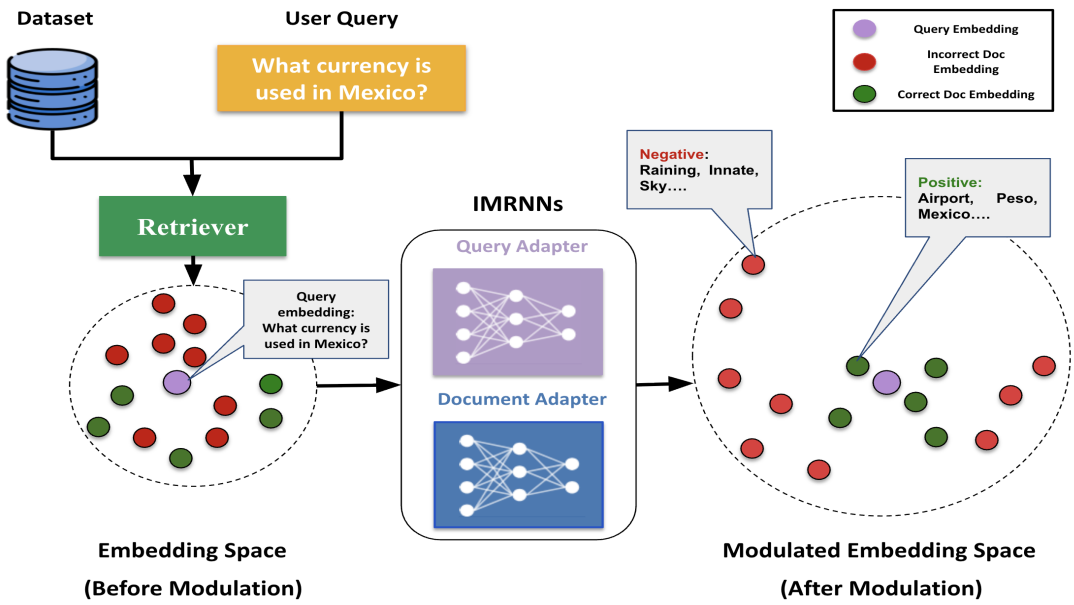

Figure 1: **An Example Showcasing `IMRNNs`'s Mechanism.** The Retriever receives the User Query and documents from the Dataset, creating a static **Embedding Space**. `IMRNNs` operates on this static space and makes it dynamic by bidirectionally modulating the query and document embeddings, creating the **Modulated Embedding Space**. This modulation brings relevant documents closer to the query embedding while pushing the embeddings of irrelevant documents farther away. The Modulated Embedding Space is interpretable by calculating the modulation vector and identifying its associated key tokens. As shown in the figure, modulation vectors for positive documents correspond to keywords like *Peso* and *Mexico*, while those for negative documents correspond to keywords like *Raining* and *Sky*.

SEARCH-ADAPTOR learns two small correction matrices, one for queries and one for documents, which are applied consistently to the whole dataset, over every example. This adjusts each query's scores, yet users cannot tell which semantic aspect the correction represents; the mapping remains a trained, opaque transformation. HYPENCODER uses a hyper network that, for each query, produces a small multilayer perception (MLP)-based scorer to rate documents, replacing plain cosine similarity with a tuneable relevance function. Although this scorer is tailored to the query, its internal weights *do not reveal* which query concepts drove the decision, so the mechanism remains opaque. DIME (Dimension Importance Estimation) measures the ℓ_2 magnitude of each embedding dimension, which represent its overall strength across all embeddings, and retains only the top k% (e.g., dimensions 1–256 out of 1024). This shows which dimensions were selected but not what they represent semantically. In summary, existing adapters operate at the level of dimensions or learned functions, but cannot translate their operations into human-understandable concepts like keywords or semantic features.

We introduce **Interpretable Modular Retrieval Neural Networks** (`IMRNNs`), the first retrieval adapter architecture designed to simultaneously improve performance and provide *multi-level interpretability*. `IMRNNs` are *inherently interpretable*, that is, no external method is required to explain/interpret the results. `IMRNNs` consists of two lightweight MLP modules operating on static embeddings: a *Query Adapter* that modulates document embeddings conditioned on the query (pulling relevant documents closer and pushing irrelevant ones away), and a *Document Adapter* that aggregates corpus-level signals to adapt the query embedding. `IMRNNs` uses bidirectional modulation mechanism to achieves three forms of interpretability *ignored* in prior work:

- *Structural interpretability*: The modulation mechanism uses explicit affine transformations $\mathbf{W}_q\mathbf{d} + \mathbf{b}_q$ and $\bar{\mathbf{W}}_d\mathbf{q} + \bar{\mathbf{b}}_d$, where weight matrices $\mathbf{W}_q$, $\bar{\mathbf{W}}_d$ and bias vectors $\mathbf{b}_q$, $\bar{\mathbf{b}}_d$ are directly observable. Unlike multi-layer neural scorers, users can inspect exactly which mathematical operations transformed the embeddings (Wang et al., 2023; Arendt et al., 2021).
- *Attribution-level interpretability*: By computing the difference $\Delta\mathbf{d} = \mathbf{d}_{\text{mod}} - \mathbf{d}_{\text{orig}}$ and $\Delta\mathbf{q} =$

$\mathbf{q}_{\text{mod}} - \mathbf{q}_{\text{orig}}$, we obtain the exact change vector induced by modulation. This reveals precisely which embedding dimensions (e.g., dimensions 23, 45, 167) increased or decreased, and by how much, enabling dimension-level attribution of retrieval decisions (Zhou et al., 2022; Zhang et al., 2022; Calderon and Reichart, 2025).

- *Semantic-level interpretability*: We back-project the change vectors $\Delta\mathbf{d}$ and $\Delta\mathbf{q}$ from the adapter's working space to the original encoder's token embedding space using the Moore-Penrose pseudoinverse (Barata and Hussein, 2012). We then compute cosine similarity between the back-projected vector and every token embedding in the vocabulary. This identifies specific tokens that align with the modulation direction, revealing *what semantic concepts* drove the retrieval decision in human-understandable terms (Rajagopal et al., 2021; Sajjad et al., 2022).

Crucially, the identified semantic concepts (tokens) are not post-hoc explanations; rather, they directly correspond to the mathematical transformations that changed the ranking. This contrasts with prior interpretability approaches (Yuksel and Kamps, 2025a; Llordes et al., 2023a) that analyze what a model did *after the fact*, rather than revealing the mechanism itself. Additionally, `IMRNNs` adapts differently for each query, and preserves the efficiency of cosine similarity scoring without introducing complex neural components. The complete end-to-end working of `IMRNNs` with an example is shown in Figure 1.

Contributions. Our work makes the following contributions:

- We propose `IMRNNs`, the **first retrieval adapter** to introduce structural, attribution-level, and semantic-level interpretability into dense retrieval systems, addressing a critical gap in making RAG pipelines trustworthy and transparent.
- We design a **bidirectional per-query modulation mechanism** where document embeddings adapt to the query context and query embeddings adapt to corpus-level signals, enabling semantic alignment without re-encoding or expensive cross-attention.
- We develop a **token-level attribution method** using Moore-Penrose back-projection that causally links embedding modulations to human-interpretable keywords, providing explanations grounded in the actual retrieval mechanism rather than post-hoc approximations.
- We demonstrate **significant performance gains** across seven diverse benchmarks, showing that interpretability and performance are not competing objectives and `IMRNNs` achieve both simultaneously with minimal computational overhead.

2 Problem Formulation

This section formally defines the dense retrieval task and introduces the `IMRNNs` architecture.

Task Definition: Given a user query q and a document corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, the goal of the dense retriever is to rank all documents in $\mathcal{D}$ by their relevance to q. A dense retrieval system consists of two components:(a) A **base encoder** $f_\theta : \mathcal{V} \to \mathbb{R}^n$ that maps text sequences (queries or documents) to fixed-dimensional embeddings, where $\mathcal{V}$ is the space of all possible text sequences and n is the embedding dimension. (b) A **similarity function** $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that computes relevance scores between query and document embeddings. Standard dense retrievers compute $\text{score}(q, d_i) = \cos\big(f_\theta(q), f_\theta(d_i)\big)$, where $\cos(\cdot,\cdot)$ denotes cosine similarity. Documents are then ranked in descending order of these scores.

Current Limitation With Static Embeddings: The embeddings $f_\theta(q)$ and $f_\theta(d_i)$ are *static*, computed independently and fixed after encoding. This means a document receives the same embedding regardless of which query it is being matched against, preventing the retriever from dynamically emphasizing query-relevant aspects of each document. Additionally, the system lacks interpretability: users cannot see which semantic dimensions drive the similarity score for a specific query-document pair.

2.1 Our approach: `IMRNNs`

`IMRNNs` address this limitation by introducing *dynamic modulation* on top of static embeddings from a frozen base encoder.

1. Dimension Reduction via Projection. To enable efficient learning and generalization, we project the high-dimensional base embeddings into a lower-dimensional working space. Let $\mathbf{P} \in \mathbb{R}^{m \times n}$ be a fixed linear projection matrix, where $m < n$ (we use $m = 256$ for $n = 1024$ in our experiments). For query q and document d_i, we

compute:

$$\mathbf{q}_{\text{orig}} = f_\theta(q) \in \mathbb{R}^n, \quad \mathbf{q}_{\text{proj}} = \mathbf{P}\mathbf{q}_{\text{orig}} \in \mathbb{R}^m$$
$$\mathbf{d}^{(i)}_{\text{orig}} = f_\theta(d_i) \in \mathbb{R}^n, \quad \mathbf{d}^{(i)}_{\text{proj}} = \mathbf{P}\mathbf{d}^{(i)}_{\text{orig}} \in \mathbb{R}^m$$

We optimize $\mathbf{P}$ jointly with the adapters (described next), while the encoder f_θ remains frozen.

2. Query Adapter ($\mathcal{A}_q$) is a lightweight neural network that uses projected query embedding to produce a weight matrix and bias vector to modulate all document embeddings: $\mathcal{A}_q(\mathbf{q}_{\text{proj}}) = \left(\mathbf{W}_q \in \mathbb{R}^{m\times m}, \mathbf{b}_q \in \mathbb{R}^m\right)$. The modulated document embedding for document d_i is then computed as: $\mathbf{d}^{(i)}_{\text{mod}} = \mathbf{W}_q \cdot \mathbf{d}^{(i)}_{\text{proj}} + \mathbf{b}_q$. This transformation allows the query to *pull* semantically relevant documents closer in the embedding space and *push* irrelevant documents farther away, adapting document representations to the specific query context.

3. Document Adapter ($\mathcal{A}_d$) is a separate second lightweight neural network that processes each document embedding independently to produce document-specific transformations: $\mathcal{A}_d(\mathbf{d}^{(i)}_{\text{proj}}) = \left(\mathbf{W}^{(i)}_d \in \mathbb{R}^{m\times m}, \mathbf{b}^{(i)}_d \in \mathbb{R}^m\right)$. These transformations are aggregated across all documents to create a corpus-level adaptation signal:

$$\bar{\mathbf{W}}_d = \frac{1}{N}\sum_{i=1}^{N} \mathbf{W}^{(i)}_d, \quad \bar{\mathbf{b}}_d = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}^{(i)}_d$$

The modulated query embedding is then computed as: $\mathbf{q}_{\text{mod}} = \bar{\mathbf{W}}_d \cdot \mathbf{q}_{\text{proj}} + \bar{\mathbf{b}}_d$. This enables the query embedding to adapt to the characteristics of the document corpus, thereby aligning with the vocabulary and semantic space of the available documents.

4. Scoring and Ranking. After bidirectional modulation, the final relevance score between query q and document d_i is computed using cosine similarity:

$$\text{score}_{\text{IMRNNs}}(q, d_i) = \cos\left(\mathbf{q}_{\text{mod}}, \mathbf{d}^{(i)}_{\text{mod}}\right)$$

Layer normalization is applied to both $\mathbf{q}_{\text{mod}}$ and $\mathbf{d}^{(i)}_{\text{mod}}$ before computing cosine similarity to ensure stable gradients and bounded scores.

2.2 Training

`IMRNNs` are trained using a margin-based ranking loss over query-document pairs. For each query q in a training batch $\mathcal{B}$, we sample one relevant document d^+ and one irrelevant document d^- with BM25 for a hard negative example. The loss function is:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{(q,d^+,d^-)\in\mathcal{B}} \max\Big\{0, \gamma - \cos(\mathbf{q}_{\text{mod}}, \mathbf{d}^+_{\text{mod}}) + \cos(\mathbf{q}_{\text{mod}}, \mathbf{d}^-_{\text{mod}})\Big\}$$

where $\gamma > 0$ is a margin hyperparameter that enforces a minimum separation between relevant and irrelevant pairs in the modulated embedding space.

During training, only the parameters of $\mathcal{A}_q$ and $\mathcal{A}_d$ are updated while the base LLM encoder, e5-large (Wang et al., 2024), f_θ remain frozen. We optimize using the Adam optimizer with weight decay regularization to prevent overfitting. To bound computational cost and memory requirements, training operates on a subset of top-k BM25-retrieved candidates per query rather than the entire corpus. Early stopping is applied based on validation set performance to determine the optimal number of training epochs. Additional hyperparameter settings and implementation details are provided in Section 3.

2.3 Inference Workflow

At inference time, `IMRNNs` operate in the following steps:

1. For a new query q, compute $\mathbf{q}_{\text{orig}} = f_\theta(q)$ and $\mathbf{q}_{\text{proj}} = \mathbf{P}\mathbf{q}_{\text{orig}}$. For all documents in the corpus (pre-computed offline), obtain $\{\mathbf{d}^{(i)}_{\text{proj}}\}_{i=1}^N$.
2. Pass $\mathbf{q}_{\text{proj}}$ through $\mathcal{A}_q$ to obtain $(\mathbf{W}_q, \mathbf{b}_q)$, then compute modulated document embeddings:
$$\mathbf{d}^{(i)}_{\text{mod}} = \mathbf{W}_q \cdot \mathbf{d}^{(i)}_{\text{proj}} + \mathbf{b}_q \quad \text{for } i = 1, \ldots, N$$
3. Pass each $\mathbf{d}^{(i)}_{\text{proj}}$ through $\mathcal{A}_d$ to obtain $\{(\mathbf{W}^{(i)}_d, \mathbf{b}^{(i)}_d)\}_{i=1}^N$, then aggregate and modulate the query:
$$\mathbf{q}_{\text{mod}} = \bar{\mathbf{W}}_d \cdot \mathbf{q}_{\text{proj}} + \bar{\mathbf{b}}_d$$
4. For each document d_i, compute $\text{score}_{\text{IMRNNs}}(q, d_i) = \cos(\mathbf{q}_{\text{mod}}, \mathbf{d}^{(i)}_{\text{mod}})$ and rank documents in descending order of score.

Computational Efficiency: The adapter networks $\mathcal{A}_q$ and $\mathcal{A}_d$ are lightweight (2-layer MLPs), adding minimal overhead. Document projections $\mathbf{d}^{(i)}_{\text{proj}}$ can be pre-computed offline and cached, so adapter forward passes dominate online cost per query and scale linearly with corpus size.

2.4 Interpretability Mechanism

The modulation framework enables direct interpretability by analyzing the change induced in

the original embedding space . For any query-document pair (q, d_i), we compute the modulation vectors:

$$\Delta\mathbf{q} = \mathbf{q}_{\text{mod}} - \mathbf{q}_{\text{proj}}$$

$$\Delta\mathbf{d}^{(i)} = \mathbf{d}^{(i)}_{\text{mod}} - \mathbf{d}^{(i)}_{\text{proj}}$$

The change in retrieval score induced by modulation is:

$$\Delta\text{similarity} = \cos(\mathbf{q}_{\text{mod}}, \mathbf{d}^{(i)}_{\text{mod}}) - \cos(\mathbf{q}_{\text{proj}}, \mathbf{d}^{(i)}_{\text{proj}})$$

Positive Δ similarity indicates the modulation pulled the query-document pair closer (increasing relevance), while negative values indicate they were pushed apart (decreasing relevance). To interpret these vectors in terms of semantic concepts, we back-project them to the original encoder's embedding space using the Moore-Penrose pseudoinverse $\mathbf{P}^+ = \mathbf{P}^\top(\mathbf{P}\mathbf{P}^\top)^{-1}$ as follows:

$$\Delta\mathbf{q}_{\text{orig}} = \mathbf{P}^+\Delta\mathbf{q}$$

$$\Delta\mathbf{d}^{(i)}_{\text{orig}} = \mathbf{P}^+\Delta\mathbf{d}^{(i)}.$$

In general, with $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (SVD), we can say that $\mathbf{P}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top$.

Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}|\times n}$ denote the encoder's token embedding table, where each row $\mathbf{e}_t$ is the embedding of token $t \in \mathcal{V}$. We compute cosine similarity between the back-projected modulation vector and each token embedding:

$$\text{score}_t = \cos(\Delta\mathbf{q}_{\text{orig}}, \mathbf{e}_t) = \frac{\Delta\mathbf{q}_{\text{orig}}^\top \mathbf{e}_t}{\|\Delta\mathbf{q}_{\text{orig}}\|_2 \|\mathbf{e}_t\|_2}$$

Tokens with high positive scores indicate concepts that the modulation emphasized (pulling the embedding toward), while tokens with high negative scores indicate concepts that were de-emphasized (pushing away). By ranking tokens by $|\text{score}_t|$ and examining the top-ranked tokens, we obtain human-interpretable explanations of what semantic features drove the retrieval decision.

3 Experimental Setup

We evaluate `IMRNNs` on two complementary tasks across diverse datasets, comparing against state-of-the-art retrieval adapter baselines.

Retrieval Task. In the retrieval task, we determine the effectiveness of `IMRNNs` by comparing document retrieval accuracy with the base dense retriever and recent popular competing adaptation methods. Performance gain is evaluated on BEIR's held-out test sets using MRR, nDCG, and Recall.

Interpretability Task. We qualitatively examine the semantic transparency of `IMRNNs`' modulation mechanism. For selected queries, we analyze modulation vectors for both relevant and irrelevant documents, computing the Δ similarity (the change from the original to the modulated scores) and back-projecting the modulation vectors to identify top-ranked tokens using the Moore-Penrose pseudoinverse (Section 2). We validate interpretability by examining whether the identified keywords align with document relevance and whether changes in ranking correlate with meaningful semantic shifts.

Dataset Selection Rationale. The seven BEIR benchmark datasets stress-test `IMRNNs` across multiple dimensions. MS MARCO tests scalability with large-scale web search (8.8M documents). Natural Questions and HotpotQA both use Wikipedia, but HotpotQA requires multi-hop reasoning to synthesize answers from multiple passages, testing whether modulation connects semantically related but lexically distinct evidence. SciFact and TREC-COVID evaluate domain-specific terminology and precise semantic matching in scientific/biomedical retrieval. FiQA-2018 tests adaptation to financial jargon and numerical reasoning. Webis-Touché 2020 involves argumentation retrieval where relevance depends on identifying viewpoints rather than topical overlap. Together, these span corpus sizes from 5K to 8.8M documents, single-hop to multi-hop reasoning, and general to highly specialized domains. Dataset statistics are in Table 5 in the Appendix.

Baselines. We compare three adapter families:

- **DIME variants**: Select embedding subspaces by ranking dimensions via ℓ_2 magnitude. We evaluate 20%, 40%, 60%, and 80% dimensionality reduction to test whether simple dimension selection competes with learned modulation.
- **SEARCH-ADAPTOR**: Learns dataset-specific residual transformations applied globally to all queries and documents. This evaluates whether static, dataset-level adaptation suffices versus `IMRNNs`' per-query modulation.
- **HYPENCODER variants**: Generates query-specific MLPs to score documents, replacing cosine similarity with learned neural functions. We evaluate 2, 4, 6, and 8 hidden layer configurations to test whether more expressive scoring functions outperform `IMRNNs`' simpler modulation while maintaining interpretability.

All baselines operate on identical frozen encoder embeddings and use the same data splits. Implementation details are provided in Appendix C.

Table 1: **Qualitative Examples of Modulation Effects on Retrieval Performance.** For each query, we show a relevant (green) and irrelevant (red) document pair. Modulation modestly reduces similarity for relevant documents (Δ Sim. ≈ -0.06) while drastically reducing it for irrelevant ones (Δ Sim. ≈ -0.30), improving discrimination. This enables relevant documents to rise in rankings (e.g., entering top-5 as shown in rightmost columns). Keywords extracted from queries and documents aid interpretability. Green indicates relevant; red indicates irrelevant.

Query	Document Type	Document Text	Orig. Sim.	Mod. Sim.	Δ Sim.	Query Keywords	Document Keywords	Orig Top-5 Doc. (Before)	Mod Top-5 Doc. (After)
What county is incline village nv?	Relevant	(Redirected from Incline Village–Crystal Bay, Nevada) Incline Village is a census–designated place in Washoe County, Nevada on the north shore …	0.91	0.85	−0.06	Incline, Rupert	Incline, Hilly, Nevada	1. No 2. No 3. No 4. No 5. No	1. No 2. No 3. No 4. No 5. **Yes**
	Irrelevant	*Public Meetings: The Alva Regional Airport Commission will …*	0.41	0.11	−0.29	Incline, Rupert	Airport, Regional, Sanctioned	–	–
What currency is used in Mexico?	Relevant	Mexico, Peso. The Mexican Peso is the currency of Mexico. Our currency rankings show that the most popular Mexico Peso exchange …	0.94	0.87	−0.06	Mexican, Ridges	Airport, Peso, Mexico	1. No 2. **Yes** 3. No 4. No 5. No	1. **Yes** 2. No 3. No 4. No 5. No
	Irrelevant	*All you have to do is tune to the right channel or visit any number of weather and news Web sites and …*	0.39	0.05	−0.34	Mexican, Ridges	Raining, Innate, Sky	–	–
Do diabetic patients with acute coronary syndrome experience decreased short-term and long-term	Relevant	**BACKGROUND** Diabetes mellitus is a major risk factor for adverse outcomes after acute coronary syndromes (ACS). Because …	0.93	0.84	−0.08	Coronary, Risk, Bleeding	Diabetes, Increased, Senses	1. No 2. No 3. No 4. No 5. No	1. No 2. No 3. **Yes** 4. No 5. No
	Irrelevant	*Global, regional, and national comparative risk assessment of 79 behavioural, environmental …*	0.53	0.19	−0.34	Coronary, Risk, Bleeding	Behavioural, Metabolism	–	–

4 Discussion on Interpretability

Interpretability Task. Table 1 shows that `IMRNNs` explain retrieval decisions through keywords that directly cause score changes. Consider the query "What currency is used in Mexico?" The similarity of the relevant document drops slightly from 0.94 to 0.87 (Δ = -0.07), and the extracted keywords are {`peso`, `Mexico`}. To understand it better, consider the above tokens: the modulation strengthened these currency concepts in the embedding, refining the match. The irrelevant document about weather websites plummets from 0.39 to 0.05 (Δ = -0.34), with keywords {`raining`, `sky`}; the modulation identified these off-topic concepts and pushed the document away. The keywords explain a shift in the embeddings, making more relevant concepts and away from less relevant ones.

This pattern holds across all queries. For "*What county is incline village nv?*", the keywords {`Incline`, `Nevada`} identify the location terms that distinguish the correct document from an airport commission document with keywords {`Airport`, `Regional`}. The score drops from 0.41 to 0.11 for the wrong document, as `IMRNNs` recognized that these terms don't answer a location query. For the medical query about diabetic bleeding risk, the correct document surfaces {`diabetes`, `coronary`} while the wrong document shows {`behavioural`, `metabolism`}.

Table 2: **Results on Open-Domain Datasets (MS MARCO, Natural Questions, and HotpotQA).**

Methods	MS MARCO			Natural Questions			HotpotQA			Average		
	nDCG	R	MRR	nDCG	R	MRR	nDCG	R	MRR	nDCG	R	MRR
e5-large-v2	0.85	0.97	0.81	0.72	0.90	0.68	0.54	0.61	0.65	0.70	0.83	0.71
DIME 20%	0.85	0.97	0.81	0.72	0.89	0.68	0.53	0.60	0.64	0.70	0.82	0.71
DIME 40%	0.84	0.97	0.80	0.72	0.90	0.67	0.51	0.58	0.62	0.69	0.82	0.70
DIME 60%	0.83	0.96	0.79	0.70	0.88	0.66	0.47	0.54	0.57	0.67	0.79	0.67
DIME 80%	0.80	0.93	0.75	0.64	0.83	0.60	0.37	0.42	0.45	0.60	0.73	0.60
SearchAd	0.84	0.96	0.80	0.71	0.89	0.67	0.37	0.43	0.46	0.64	0.76	0.64
Hyp (2)	0.68	0.74	0.71	0.43	0.60	0.40	0.29	0.25	0.27	0.47	0.53	0.46
Hyp (4)	0.69	0.74	0.70	0.44	0.60	0.40	0.29	0.26	0.27	0.47	0.53	0.46
Hyp (6)	0.71	0.75	0.71	0.44	0.61	0.40	0.30	0.26	0.28	0.48	0.54	0.46
Hyp (8)	0.70	0.75	0.72	0.45	0.61	0.42	0.30	0.26	0.29	0.48	0.54	0.48
IMRNNs	**0.88**	**0.99**	**0.85**	**0.75**	**0.93**	**0.71**	**0.59**	**0.63**	**0.66**	**0.74**	**0.85**	**0.74**

The IMRNNs identify which terms are relevant to the specific medical question being asked.

The Moore-Penrose pseudoinverse makes this interpretability possible by mapping modulation vectors back to actual vocabulary tokens. Without it, we would only see that scores changed, not which concepts drove those changes. The method identifies tokens whose embeddings align with the modulation direction; these are the semantic features that the adapters emphasized or suppressed. peso appears in the keyword list, clearly indicating that the modulation moved the embedding toward the dimension where peso lives in the vocabulary space.

The ranking changes validate that keywords capture real semantic reasoning. In queries 1 and 3, the correct document doesn't appear in the original top-5 but enters after modulation. Here the extracted keywords, {Incline, Nevada} and {diabetes, coronary}, are the key concepts. In query 2, the correct document advances from rank 2 to rank 1, with peso as the decisive keyword. The system isn't just rescoring randomly as it identifies relevant semantic concepts for a given query and adjusts the embeddings accordingly.

Relevant documents show small similarity changes (Δ = -0.06 to -0.08) because the modulation preserves already-good matches while refining them. Irrelevant documents show large drops (Δ = -0.29 to -0.34) because the modulation actively suppresses mismatches. This asymmetry demonstrates principled behavior: it strengthens correct alignments and weakens incorrect ones.

5 Discussion on Retrieval Performance

Tables 2, 3, and 4 demonstrate that IMRNNs' bidirectional modulation mechanism consistently improves ranking quality across diverse retrieval scenarios. The performance gains stem from how the adapters reshape the embedding space: the Query Adapter generates transformations that pull semantically relevant documents closer to the query while pushing irrelevant ones away, and the Document Adapter aggregates corpus-level signals to help the query align with the vocabulary and semantic structure of available documents. This dual adaptation addresses the core limitation of static embeddings, they cannot adjust to query-specific relevance signals or corpus-specific terminology. The nDCG improvements are particularly meaningful because this metric rewards systems that place highly relevant documents at top positions with logarithmic discounting (Järvelin and Kekäläinen, 2002). By modulating embeddings to strengthen correct semantic alignments, IMRNNs ensures that the most relevant documents rise to ranks 1-3 where users actually look, rather than languishing at ranks 8-10 where they contribute little to user satisfaction.

The baseline methods' performance reveals why per-query semantic adaptation matters. SEARCH-ADAPTOR learns dataset-level residual transformations that apply uniformly across all queries, but struggles because different queries require different semantic adjustments, a geographic query about "incline village" needs location-term emphasis, while a medical query about "diabetic bleeding risk" needs disease-term emphasis. A single global transformation cannot capture this diversity, explaining why SEARCH-ADAPTOR often underperforms the base retriever. DIME variants progressively degrade as they remove more dimensions, demonstrating that magnitude-based selection discards information critical for semantic matching, a dimension that appears unimportant globally may be essential for specific queries. HYPENCODER generates

Table 3: **Results on Domain-Specific Datasets (Scifact, Trec-COVID, and Webis-Touche2020).**

Methods	Scifact			Trec-COVID			Webis-Touche			Average		
	nDCG	R	MRR	nDCG	R	MRR	nDCG	R	MRR	nDCG	R	MRR
e5-large-v2	0.72	0.87	0.68	0.79	0.03	**1.00**	0.60	0.31	0.85	0.70	0.40	0.84
DIME 20%	0.72	0.87	0.67	0.82	0.03	**1.00**	0.59	0.30	0.85	0.71	0.40	0.84
DIME 40%	0.70	0.87	0.65	0.82	0.03	**1.00**	0.61	0.32	0.85	0.71	0.41	0.83
DIME 60%	0.68	0.84	0.64	0.84	0.03	0.94	0.60	0.31	0.85	0.71	0.39	0.81
DIME 80%	0.60	0.78	0.55	0.77	0.03	0.86	**0.62**	0.32	0.78	0.66	0.38	0.73
SearchAd	0.69	0.85	0.65	0.84	0.03	**1.00**	0.58	0.31	0.85	0.70	0.40	0.83
Hyp (2)	0.62	0.75	0.63	0.78	0.02	0.59	0.59	0.29	0.83	0.66	0.35	0.68
Hyp (4)	0.61	0.63	0.68	0.76	0.01	0.44	0.59	0.28	0.83	0.65	0.31	0.65
Hyp (6)	0.72	0.62	0.60	0.79	0.01	0.47	0.60	0.29	0.84	0.70	0.31	0.64
Hyp (8)	0.65	0.81	0.68	0.84	0.01	0.52	0.60	0.29	0.85	0.70	0.37	0.68
IMRNNs	**0.74**	**0.88**	**0.70**	**0.85**	**0.04**	**1.00**	**0.62**	**0.34**	**0.91**	**0.74**	**0.42**	**0.87**

Table 4: **Results on FiQA.**

Methods	FiQA		
	nDCG	R	MRR
e5-large-v2	0.20	0.23	0.28
DIME 20%	0.20	0.22	0.28
DIME 40%	0.19	0.21	0.27
DIME 60%	0.18	0.20	0.25
DIME 80%	0.14	0.16	0.23
SearchAd	0.17	0.20	0.25
Hyp (2)	0.32	0.40	0.39
Hyp (4)	0.31	0.42	0.35
Hyp (6)	0.32	0.42	0.37
Hyp (8)	0.33	0.42	0.39
IMRNNs	0.22	0.24	0.29
IMRNNs (Hyp(8))	**0.36**	**0.44**	**0.42**

query-conditioned neural scorers, but these black-box functions lack the explicit semantic grounding that IMRNNs' modulation vectors provide, and adding more layers (2→8) yields diminishing returns without addressing the fundamental need for interpretable semantic adaptation. The consistent pattern across open-domain datasets (MS MARCO, Natural Questions, HotpotQA) and specialized domains (SciFact, TREC-COVID, FiQA, Webis-Touché) confirms that IMRNNs' approach generalizes: the mechanism adapts to whatever semantic features matter for each dataset, whether lexical overlap, multi-hop reasoning, domain terminology, or argumentative stance.

The FiQA anomaly (Table 4) illuminates a critical dependency: IMRNNs amplifies the quality of their base embeddings rather than replacing them. When e5-large-v2 embeddings are poorly calibrated for financial terminologies, IMRNNs built atop them cannot match HYPENCODER, which uses its own embedding generation. However, stacking IMRNNs on top of Hypencoder (8) embeddings yields substantial gains (9.09% nDCG, 4.76% Recall, 7.68% MRR), demonstrating that the modulation mechanism successfully enhances any sufficiently rich embedding space. This reveals IMRNNs 's architectural advantage: the adapters operate as a plug-and-play layer that improves whatever base retriever provides the best embeddings for a given domain, rather than requiring full model retraining or domain-specific architecture changes. The modulation vectors capture semantic refinements that static embeddings miss, emphasizing currency-related dimensions for financial queries, location dimensions for geographic queries, and disease dimensions for medical queries, while preserving the computational efficiency of cosine similarity.

6 Conclusion

We introduce IMRNNs, the first lightweight retrieval adapters that make the embeddings of dense retrievers interpretable by achieving three levels of interpretability: structural, attribution, and semantic. We benchmark the retrieval performance of IMRNNs against state-of-the-art retrieval adapter baselines on diverse datasets and demonstrate that IMRNNs adapts query and document embeddings more effectively than competing methods, while also showing strong generalization. The semantic-level interpretability of IMRNNs is especially useful in applications where access to key tokens or keywords in both the query and the documents plays a major role. We will release all code and experimental scripts for reproducibility.[1]

[1] https://anonymous.4open.science/r/IMRNNs-5F23/README.md

7 Limitations

Experiments across diverse datasets reveal three limitations of `IMRNNs`. First, its performance depends on the semantic richness of the embedding space produced by the base dense retriever. Second, there is currently no reliable way to predict how much improvement `IMRNNs` will achieve on a particular dataset without running experiments on that dataset. Third, some key tokens identified from the resultant modulation vectors may appear random or not meaningful, as mapping an embedding to its exact semantic token is extremely difficult and removing such tokens is often necessary to obtain a clear, interpretable picture.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31.

Dustin Arendt, Zhuanyi Shaw, Prasha Shrestha, Ellyn Ayton, Maria Glenski, and Svitlana Volkova. 2021. Crosscheck: Rapid, reproducible, and interpretable model evaluation. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 79–85.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

João Carlos Alves Barata and Mahir Saleh Hussein. 2012. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165.

Nitay Calderon and Roi Reichart. 2025. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693.

Cesare Campagnano, Antonio Mallia, and Fabrizio Silvestri. 2025a. Unveiling dime: Reproducibility, generalizability, and formal analysis of dimension importance estimation for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3367–3376.

Cesare Campagnano, Antonio Mallia, and Fabrizio Silvestri. 2025b. Unveiling dime: Reproducibility, generalizability, and formal analysis of dimension importance estimation for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3367–3376, New York, NY, USA. Association for Computing Machinery.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: sparse lexical and expansion model for first stage ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2288–2292. ACM.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Hao Kang, Tevin Wang, and Chenyan Xiong. 2025. Interpret and control dense retrieval with sparse latent features. In *Proceedings of the 2025 Conference of*

the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 700–709, Albuquerque, New Mexico. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Anirudh Khatry, Yasharth Bajpai, Priyanshu Gupta, Sumit Gulwani, and Ashish Tiwari. 2023. Augmented embeddings for custom retrievals. *Preprint*, arXiv:2310.05380.

Julian Killingback, Hansi Zeng, and Hamed Zamani. 2025a. Hypencoder: Hypernetworks for information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2372–2383.

Julian Killingback, Hansi Zeng, and Hamed Zamani. 2025b. Hypencoder: Hypernetworks for information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2372–2383, New York, NY, USA. Association for Computing Machinery.

Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. 2025. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems. *arXiv preprint arXiv:2503.15191*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 287–296.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Michael Llordes, Debasis Ganguly, Sumit Bhatia, and Chirag Agarwal. 2023a. Explain like i am bm25: Interpreting a dense model's ranked-list with a sparse approximation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1976–1980.

Michael Llordes, Debasis Ganguly, Sumit Bhatia, and Chirag Agarwal. 2023b. Explain like i am bm25: Interpreting a dense model's ranked-list with a sparse approximation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1976–1980, New York, NY, USA. Association for Computing Machinery.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi A Omiye, Mehr Kashyap, and 1 others. 2025. Factehr: A dataset for evaluating factuality in clinical notes using llms.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850.

Stephen Robertson and Hugo Zaragoza. 2009a. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Stephen Robertson and Hugo Zaragoza. 2009b. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Nicholas Rossi, Juexin Lin, Feng Liu, Zhen Yang, Tony Lee, Alessandro Magnani, and Ciya Liao. 2024. Relevance filtering for embedding-based retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4828–4835, New York, NY, USA. Association for Computing Machinery.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. 2024a. Search-adaptor: Embedding customization for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12230–12247.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. 2024b. Search-adaptor: Embedding customization for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12230–12247, Bangkok, Thailand. Association for Computational Linguistics.

Goksenin Yuksel and Jaap Kamps. 2025a. Interpretability analysis of domain adapted dense retrievers. *arXiv preprint arXiv:2501.14459*.

Goksenin Yuksel and Jaap Kamps. 2025b. Interpretability analysis of domain adapted dense retrievers. *Preprint*, arXiv:2501.14459.

Sepanta Zeighami, Zac Wellmer, and Aditya Parameswaran. 2025. NUDGE: Lightweight non-parametric fine-tuning of embeddings for retrieval. In *The Thirteenth International Conference on Learning Representations*.

Sheng Zhang, Jin Wang, Haitao Jiang, and Rui Song. 2022. Locally aggregated feature attribution on natural language model understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2189–2201.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2022. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

A Related Work

We organize related work into three categories corresponding to the core components IMRNNs build upon: the dense retrieval foundation, the adapter mechanisms for performance enhancement, and the interpretability gap we address.

Dense Retrieval Methods. RAG systems typically employ a two-stage pipeline: an initial retriever selects candidate documents efficiently from large corpora, followed by a reranker that refines the ranking. Initial retrievers span lexical methods like BM25 (Robertson and Zaragoza, 2009b), which provide transparent term-matching but struggle with semantic gaps; dense bi-encoders like DPR (Karpukhin et al., 2020) and Contriever (Izacard et al., 2022), which capture semantic similarity through learned embeddings but operate as black boxes; and hybrid approaches like SPLADE (Formal et al., 2021), which combine contextual encoders with sparse representations. We focus on dense bi-encoders because they dominate modern RAG systems due to their favorable accuracy-efficiency tradeoff for initial retrieval over large corpora, and improving their interpretability addresses a critical gap that lexical methods solve trivially but dense methods completely lack.

Retrieval Adapters. Recent work has introduced lightweight adapter modules that enhance dense retrievers without full model retraining. We categorize these by operational scope to clarify where IMRNNs fit architecturally and to show that interpretability remains absent across all adapter types. *Architecture-modifying adapters* (Rossi et al., 2024; Zeighami et al., 2025; Ding et al., 2023) update internal retriever components or scoring mechanisms, typically requiring white-box access to model internals. *Embedding-space adapters* operate on frozen encoder outputs and are most relevant to our work as they share the same operational paradigm: SEARCH-ADAPTOR (Yoon et al., 2024b), DIME (Campagnano et al., 2025b), and HYPENCODER (Killingback et al., 2025b) serve as our primary baselines because they represent the state-of-the-art in embedding-space adaptation with diverse strategies (detailed technical comparison and experimental setup in Section 3). *Parameter-efficient fine-tuning approaches* modify encoder parameters minimally: TART (Asai et al., 2023) and INSTRUCTOR (Su et al., 2023) inject task-specific instructions; LoRA (Hu et al., 2022) and IA3 (Liu et al., 2022) update low-rank subspaces; PROMPTAGATOR (Dai et al., 2023) synthesizes training data via LLMs. We exclude these from our experimental comparison because they require encoder fine-tuning (violating the frozen-encoder constraint that enables IMRNNs' plug-and-play deployment) and still provide no interpretability mechanisms. ADAPTED DENSE RETRIEVAL (Khatry et al., 2023) learns low-rank residuals for heterogeneous retrieval settings. Adapters have also been explored for sparse retrievers and rerankers (Hu et al., 2023), which fall outside our focus on dense initial retrieval. While these methods improve retrieval accuracy through various adaptation mechanisms, **none provide interpretability**: users cannot determine which semantic features drove retrieval decisions or how queries and documents are semantically aligned during ranking.

Interpretability in Information Retrieval. While interpretability has been explored in other RAG components, the initial dense retriever remains largely unaddressed. For lexical retrievers, term-matching provides inherent transparency, but dense retrievers lack comparable explanations. We examine existing interpretability approaches for dense retrieval in detail because understanding why they fall short is essential to justify IMRNNs' novel contribution. These approaches fall into three categories, each with critical limitations that IMRNNs address:

- *Surrogate approximations* (Llordes et al., 2023b) fit sparse models to approximate dense rankings, extracting per-query term weights. However, these explanations are post-hoc reconstructions that may not faithfully represent the actual dense model's decision process, and the approximation quality degrades as the gap between sparse and dense representations widens.
- *Gradient-based attribution* (Yuksel and Kamps, 2025b) applies techniques like Integrated Gradients to bi-encoders, identifying which input tokens receive high gradient signals during domain adaptation. While these methods show which tokens the model attended to during training, they do not reveal which semantic concepts the model emphasized in the embedding space during inference, and gradient-based explanations can be unstable or misleading (Adebayo et al., 2018).
- *Concept mapping* (Kang et al., 2025) aligns latent embedding dimensions with human-interpretable descriptors using sparse probing. This provides dimension-level interpretability but requires ad-

ditional annotation effort and does not directly explain how specific query-document pairs interact during retrieval.

B BEIR Statistics

Table 5: **BEIR Dataset Details.**

Dataset	Domain	Type	#Queries	#Documents
MS MARCO	Web Search	Single-hop	6,980	8,841,823
Natural Questions	Wikipedia	Single-hop	3,452	2,681,468
HotpotQA	Wikipedia	Multi-hop	7,405	5,233,329
SciFact	Scientific	Single-hop	300	5,183
TREC-COVID	Biomedical	Single-hop	50	171,332
Webis-Touché 2020	Argumentation	Single-hop	49	382,545
FiQA-2018	Finance	Single-hop	648	57,638

C Implementation Details

IMRNNs use projection dimension $m = 256$. Both adapters are two-layer MLPs with ReLU activations and layer normalization. Training uses Adam (learning rate 10^{-4}, weight decay 10^{-5}, batch size 32) with margin $\gamma = 0.3$, operating on top-100 BM25 candidates per query. Early stopping uses patience of 5 epochs on validation nDCG. Training converges within 10-20 epochs on a single NVIDIA H100 GPU.

D Computational Time Comparison

Latency is averaged per single query and throughput is computed as inverse of average latency.

Table 6: Inference efficiency comparison (averaged over 1,000 queries).

Method	Latency (ms/query)	Throughput (queries/s)
Magnitude DIME	0.96	61.22
Search-Adaptor	1.02	58.82
Hypencoder	1.71	35.09
IMRNNs	1.64	36.59