

Beyond Single Vector: Query-Aware Multi-Perspective Document Embeddings for Retrieval

Zhen Li¹, Chongyang Tao^{2*}, Yihan Chen¹, Huishuai Zhang^{1,3}, Dongyan Zhao^{1,3*}

¹Wangxuan Institute of Computer Technology, Peking University

²SKLCCSE Lab, Beihang University

³Key Laboratory of Intelligent Press Media Technology

{lizhen63, zhanghuishuai, zhaody}@pku.edu.cn, chongyangtao@gmail.com,
tash_qiuqiu@qq.com

Abstract

Recently, large language models (LLMs) have demonstrated excellent performance on general information retrieval tasks, which encode query and document into an embedding vector representing semantic information respectively. However, single-vector representations often fail to capture the multifaceted semantics of informative documents, limiting the adaptability across diverse scenarios. To address this, we propose **PRISM** (multi-PeRspective Semantic inforMation retrieval), a novel framework that leverages LLMs’ implicit reasoning capabilities to extract and structure diverse semantic perspectives. **PRISM** introduces query-aware semantic fusion to dynamically integrate relevant perspectives during retrieval through weighted query embedding interactions. Furthermore, we propose the contrastive semantic diversification, a training mechanism that disperses semantic embeddings to maximize perspective diversity. Evaluated on the English retrieval subsets of massive multilingual text embedding benchmark (MMTEB), **PRISM** achieves state-of-the-art general results compared with models trained exclusively on public data.

1 Introduction

Information retrieval has played a pivotal role in real-world applications such as web search (Nguyen et al., 2016), question answering (Chen and Yih, 2020) and retrieval-augmented generation (RAG) system (Zhao et al., 2024a). Over the past decade, dense retrieval has emerged as a key paradigm in information retrieval with the advancement of pre-trained models (Karpukhin et al., 2020; Xiong et al., 2020; Zhao et al., 2024b). A typical dense retrieval system adopts a bi-encoder architecture to encode each query and document into a single representation embedding independently. Relevance is then estimated by a similarity

measure (e.g. cosine similarity) between these embeddings, enabling efficient first-stage retrieval in information retrieval systems. This paradigm, exemplified by models like DPR (Karpukhin et al., 2020) and various BERT-based retrievers, has demonstrated substantial gains over traditional term-based methods such as TF-IDF and BM25 in capturing semantic meaning.

More recently, the emergence of large language models (LLMs) has further transformed the retrieval landscape. LLMs possess strong reasoning abilities and an extensive knowledge base, making them powerful embedding generators (Lei et al., 2024; Zhang et al., 2024; Li et al., 2024; Zhuang et al., 2024). To further tailor these embeddings to retrieval, some works fine-tune LLM encoders with contrastive objectives on query–document pairs, yielding stronger alignment and discrimination (Wang et al., 2024; Li et al., 2025; Ji et al., 2025). Based on these, progress has been made along multiple dimensions, such as enhancing training data (Wang et al., 2024; Zhang et al., 2025b), optimizing model structure (Lee et al., 2025; Muenighoff et al., 2024), and improving training objectives (BehnamGhader et al., 2024; Lee et al., 2024). Together, these approaches have achieved impressive empirical results and driven rapid adoption of LLM-based retrievers in both academic research and industrial applications.

However, despite growing progress, a key limitation remains: the single-vector paradigm. Most LLM-based retrievers still reduce queries and documents into one embedding vector each. This paradigm imposes theoretical limitations (Weller et al., 2025), and constrains a retriever to encode documents, where a single vector struggles to capture the full range of semantics in information-rich documents, especially those that embody multiple perspectives relevant to different user queries. As shown in Figure 1, documents often encompass heterogeneous facets such as factual content,

*Corresponding author

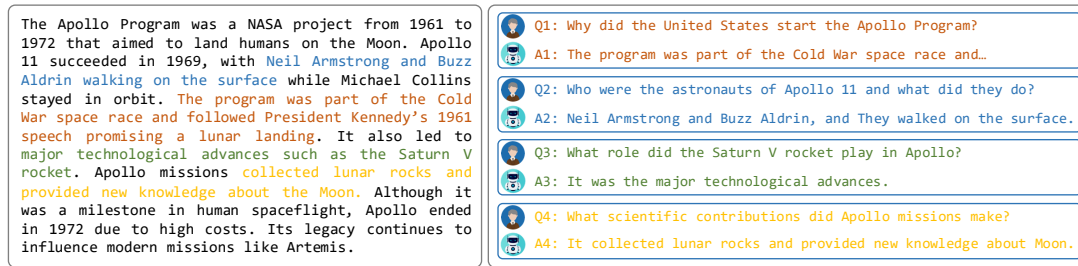


Figure 1: An example demonstrating the multi-perspective semantic document. Documents may contain diverse semantics and correspond to multiple different queries.

contextual background and so on, that cannot be faithfully preserved in a single representation. As a consequence, even strong LLM-based retrievers can under-retrieve relevant documents when a single embedding fails to capture their full semantics, leading to query–document mismatches.

Some studies have explored multi-vector or multi-view representations (Khatab and Zaharia, 2020; Gao et al., 2021a; Liao et al., 2023). These approaches generate multiple embeddings for text either by simply adding additional special tokens (Zhang et al., 2022) or using token-level embedding directly (Khatab and Zaharia, 2020), thereby capturing richer semantics and enabling fine-grained query–document matching. Although these models improve over single-vector retrievers, they are generally built upon BERT-like discriminative encoders. Their mechanisms for capturing diversity are shallow (e.g., token-level pooling) and insufficient for exploiting reasoning potential of LLMs. As a result, their ability to model multi-perspective semantics remains limited.

To resolve the above challenge, in this paper, we propose **PRISM** (Multi-PeRspective Semantic inforMation Retrieval), a novel LLM-based retrieval framework designed to overcome the bottlenecks of single-vector encoders. PRISM explicitly models documents as a set of multi-perspective embeddings, each intended to capture a distinct semantic facet. To construct these representations, PRISM leverages LLMs’ reasoning abilities, where instead of relying on surface-level token augmentations, we instruct the LLM to infer and encode multiple perspectives of a document. During retrieval, a query interacts with these perspective vectors and the resulting evidence is aggregated in a query-aware fusion strategy, ensuring that the perspectives relevant to the query dominate the relevance score. To enhance representation diversity, we introduce contrastive semantic diversification,

a training objective that disperses a document’s perspective embeddings across complementary regions of the semantic space.

Experiment results on the English retrieval subsets of massive multilingual text embedding benchmark (MMTEB) (Enevoldsen et al., 2025) demonstrate our model achieves state-of-the-art compared with models trained exclusively on public data. Extensive analysis further verifies the effectiveness of the different strategies we proposed. Our contributions are threefold:

- We propose **PRISM**, a LLM-based retriever that captures multi-perspective semantic of document to enhance the document representation embeddings leveraging the reasoning capabilities of LLM, and addresses the inherent capacity limits of single-vector encoders.
- We design a query-aware fusion mechanism and contrastive semantic diversification objective, which utilizes multiple semantic views more robustly and further optimizes representation diversity of different semantic views.
- We conduct extensive experiments and analysis on widely used benchmarks to demonstrate the effectiveness of our proposed model.

2 Related Works

2.1 Multi-vector Retrieval

Multi-vector retrieval aims to preserve richer semantics by representing texts with multiple embeddings (Humeau et al., 2020; Liao et al., 2023). ColBERT (Khatab and Zaharia, 2020) introduced the late interaction paradigm, where queries and documents are decomposed into token-level embeddings, and relevance is computed through a sum of maximum token similarities. This design maintains fine-grained semantic alignment.

Building on this, subsequent work explored efficiency and structural constraints. COIL (Gao et al., 2021a) restricts interactions to exact lexical matches, drastically reducing computation while retaining contextual embeddings. MVR (Zhang et al., 2022) explores to explicitly construct multi-view document representations, which generates several semantic views of a document and measures query–document similarity by taking the maximum similarity across all query–view pairs. These developments demonstrate a sustained effort to combine expressiveness of multi-vector representations with the efficiency required for information retrieval.

2.2 LLM-based Retrieval

With the development of pre-trained models (PTMs), information retrieval based on PTMs has gradually improved (Yu et al., 2022; Li et al., 2023b). Before the era of LLMs, numerous works leveraged PTMs such as BERT (Devlin et al., 2019) as backbones to implement retrieval methods. Recently, the emergence of LLMs has further expanded this paradigm, enabling retrieval systems to directly exploit extensive world knowledge of LLMs. Early explorations used decoder-only LLMs as embedding generators, typically representing text by the last token’s hidden state or pooling operations (Muennighoff, 2022; Su et al., 2023; Wang et al., 2024; Tao et al., 2024).

Further, some works adapt advanced strategies to improve performance (BehnamGhader et al., 2024; Lee et al., 2025; Lei et al., 2025; Yan et al., 2025; Ji et al., 2025). BGE-en-ICL (Li et al., 2025) uses in-context-learning strategy, which adds task-related few-shot examples to query during encoding, to produce high-quality text embeddings. Qwen3-embedding (Zhang et al., 2025b) designs effective data synthesis strategies, synthesizes large amounts of data and train model embedding capabilities through multi-stage optimizing.

LLM-based retrieval models have achieved strong performance across various tasks. However, existing methods mostly focus on utilizing a single embedding to represent semantically informative documents, and cannot fully utilize the powerful reasoning capabilities of LLMs. To this end, our method instructs LLM to infer the multi-perspective semantics of documents and designs query-aware fusion and contrastive semantic diversification strategies to improve the effectiveness of document semantic representation.

3 Methodology

3.1 Preliminaries of LLM-based Dense Retriever

Task Definition Given a query q and a set of candidate documents $D = \{d_1, d_2, \dots, d_N\}$, the goal of a retrieval system is to identify the subset D_q that is most relevant to q . Modern dense retrieval approaches typically follow a bi-encoder paradigm (Karpukhin et al., 2020), in which the query and each document are independently mapped into a shared latent semantic space. Formally, the bi-encoder retriever generates embeddings $E(q)$ and $E(d)$. The relevance score between query q and document d is then computed as $f(q, d) = \langle E(q), E(d) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes a lightweight similarity metric, such as the dot product or cosine similarity (Qu et al., 2021). Then relevant documents for q are retrieved: $D_q = \text{Top}K_{d \in D} f(q, d)$.

LLM-based Embedding Generation Dense retrievers based on LLMs generally derive representations from the final hidden states of LLMs, followed by various pooling strategies including special-token pooling strategy (e.g., $\langle \text{eos} \rangle$) (Li et al., 2025), mean pooling strategy averages (Lee et al., 2024), and attention-weighted pooling strategy (Wang et al., 2024; Lee et al., 2025). These mechanisms yield a single-vector representation for each query and document.

3.2 Multi-perspective LLM-based Retriever

As mentioned in previous, LLM-based retrievers typically encode text into a single embedding, which struggles to preserve the heterogeneous semantics of information rich documents. To address this limitation, we propose **PRISM**, a retrieval framework that explicitly models documents as a set of multi-perspective embeddings. Meanwhile **PRISM** integrates a query-aware fusion strategy and a contrastive semantic diversification objective to adaptively align queries with documents while preserving semantic diversity. Figure 2 demonstrates the overview of **PRISM**.

LLM-Instructed Reasoning Unlike prior multiple vector retrieval approaches that rely on shallow mechanisms such as token-level pooling, **PRISM** leverages the reasoning capabilities of LLMs to encode diverse semantic views of a document. Specifically, we instruct LLM to generate multiple perspective embeddings explicitly, via append-

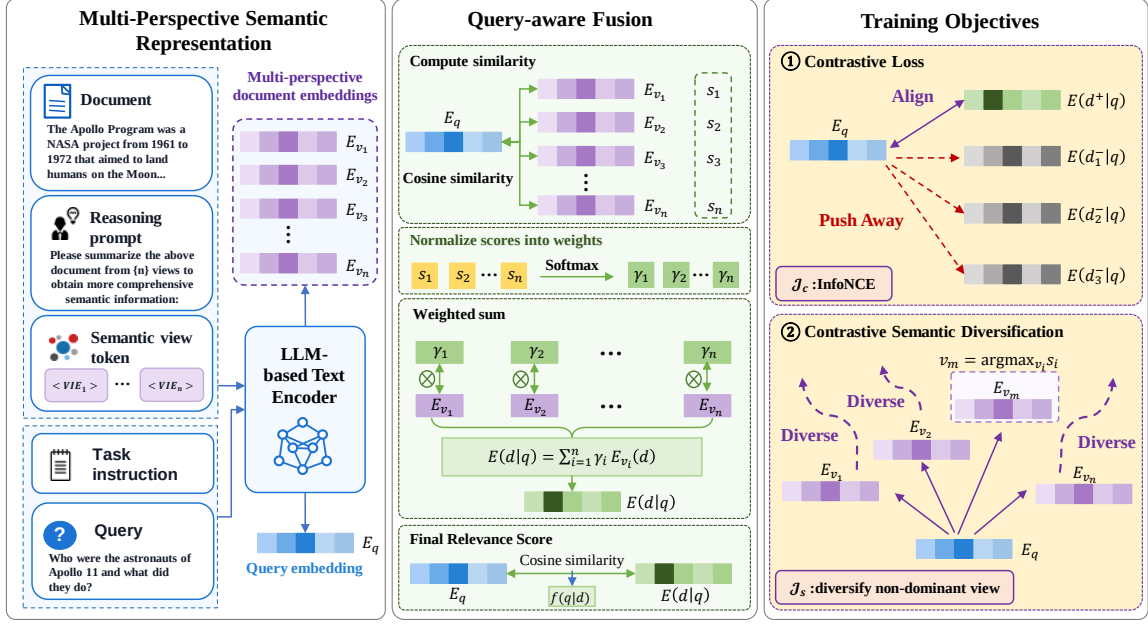


Figure 2: The overview of our proposed multi-perspective LLM-based retriever.

ing an instruction to the document which encourages LLM to reason multiple document semantic perspectives. Further, this instruction is followed by the addition of special semantic view tokens, denoted as $\{\langle \text{VIE}_1 \rangle, \langle \text{VIE}_2 \rangle, \dots, \langle \text{VIE}_n \rangle\}$, where n represents the number of semantic perspectives, and each special token serves as an anchor for one semantic perspective. The prompt we use for encoding is:

```
{Document}
Please summarize the above document from {n}
views to obtain more comprehensive semantic
information:
{⟨VIE1⟩, ⟨VIE2⟩, ..., ⟨VIEn⟩}
```

Each perspective’s embedding is derived from the last layer’s embedding of these special tokens, enabling the LLM to produce separate representations for each perspective. Formally, we denote the encoding process as:

$$E_{v_i}(d) = \text{LLM}(\{\text{document}\}; \{\text{instruction}\}; \langle \text{VIE}_1 \rangle, \dots, \langle \text{VIE}_n \rangle)[- (n - i)] \quad (1)$$

For the query, similar as the previous works (Li et al., 2025; Ji et al., 2025), each query is encoded following task-specific instructions, with the final $\langle \text{eos} \rangle$ token providing the query representation:

$$E(q) = \text{LLM}(\{\text{task_definition}\}; \{\text{query}\}; \langle \text{eos} \rangle)[-1] \quad (2)$$

Query-aware Fusion Given the multiple perspective embeddings for each document, it is necessary for retriever to design an effective late interaction strategy between the query and these embeddings for matching. To this end, we introduce a query-aware fusion strategy to aggregate the various document perspectives, which applies an attention mechanism to calculate query-aware weights for each document perspective embedding.

Specifically, we first calculate the cosine similarity between $E(q)$ and each $E_{v_i}(d)$ as the relevance score $s_i = \cos(E(q), E_{v_i}(d))$, where these scores reflect the degree to which each perspective aligns with the query. We then normalize these scores with a softmax to obtain attention weights: $\gamma_i = \text{softmax}(s_i / \tau)$.

The query-attentive document representation embedding is computed as a weighted sum of perspective embeddings:

$$E(d|q) = \sum_{i=1}^n \gamma_i E_{v_i}(d), \quad (3)$$

This formulation ensures that the final representation $E(d|q)$ dynamically emphasizes the views most pertinent to the query, effectively serving as a semantic projection of the document into the query’s context. Finally, the relevance score between query and document is defined as $f(q, d) = \cos(E(q), E(d|q))$.

The proposed query-aware fusion strategy mitigates the risk of over-relying on a single dominant

view, and enhances generalization across diverse retrieval scenarios.

Contrastive Semantic Diversification To align queries with relevant documents, following standard practice in dense retrieval, we begin with the InfoNCE (Oord et al., 2018) loss to optimize LLM-based retriever, which is widely used in contrastive learning:

$$\mathcal{J}_c = -\log \frac{\exp(f(q, d^+))}{\exp(f(q, d^+) + \sum_{d^-} \exp(f(q, d^-))} \quad (4)$$

where d^+ represents the positive documents for query, and d^- denotes negative documents which is typically sampled from both in batch negatives and hard negatives in training. This objective encourages query representation to be closer to its corresponding positive document than to negatives, thus driving effective query–document alignment.

However, despite the query-aware fusion enables effective aggregation of multiple document perspectives, the training process still faces a crucial challenge: semantic collapse. Without explicit constraints, the LLM may generate view embeddings that are highly redundant, causing different perspective embeddings to converge toward similar semantic regions. Such collapse undermines the benefit of constructing multi-perspective embeddings, as different perspectives fail to capture genuinely distinct aspects of a document.

To enforce diversification, inspired by (Zhang et al., 2022), we introduce an additional contrastive semantic diversification objective that explicitly regularizes the semantic views to specialize into different semantic regions rather than collapsing. Concretely, for each positive document d^+ , we first identify the most relevant view v_m with respect to the query q : $v_m = \arg \max_{v_i} s(q, d_{v_i})$.

This maximum value view v_m is treated as the primary alignment target. The remaining views $\{v_j\}_{j \neq m}$ are then contrasted against v_m , ensuring that they capture semantics distinct from the dominant perspective, which encourages different view embeddings to occupy complementary regions of the semantic space, thus covering diverse query intents. We define the diversification loss as:

$$\mathcal{J}_s = \log \frac{\exp(s(q, d_{v_m}^+))}{\sum_{j \in [1, n]} \exp(s(q, d_{v_j}^+))} \quad (5)$$

By applying $\mathcal{J}_s$ only to positive documents, we guarantee that there exists semantically meaning-

ful contrasted view, avoiding instability caused by irrelevant negatives.

Finally, the overall training loss combines standard InfoNCE loss with diversification term:

$$\mathcal{J}_{\text{final}} = \mathcal{J}_c + \alpha \mathcal{J}_s, \quad (6)$$

where α is a hyperparameter controlling the relative weight of diversification. Intuitively, $\mathcal{J}_c$ ensures that queries and positive documents are aligned in latent space, while $\mathcal{J}_s$ prevents the multiple views of a document from collapsing, thereby preserving semantic richness. As a result, **PRISM** learns representations that are both semantically robust and sufficiently diverse to support retrieval across heterogeneous user queries.

4 Experiments

4.1 Experiment Setup

Training We use Mistral 7B (Jiang et al., 2023) as backbone model and train **PRISM** on the publicly available training data provided by BGE-en-ICL. More details on training data are available in Appendix A. For implementation, similar as BGE-en-ICL, we finetune backbone for one epoch where we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022), with the LoRA rank to 64 and the LoRA alpha to 32, and the learning rate is set as 1e-4. Each training sample contains a query, a positive document and 7 hard negative documents. For retrieval task in training data, the batch size is set as 512 and for others, the batch size is set as 256. Meanwhile, for retrieval task, we utilize in-batch negative strategy and employ KL-divergence loss to distill knowledge from ranker following the setting of BGE-en-ICL (Li et al., 2025). We set the maximum length for encoding text as 512. We train our model on zero-shot without in context learning samples, which is different from BGE-en-ICL. The temperature τ in softmax for query-aware fusion is set as 1.0. The number of semantic perspectives is set as 4, and the hyperparameter α for contrastive semantic diversification loss is set as 0.01.

Evaluation We evaluate retrieval performance of our **PRISM** on Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025), which is an optimized version of MTEB (Muennighoff et al., 2023). We use the retrieval subtasks of MMTEB (eng, v2) for our evaluation, and we report the NDCG@10 for all benchmarks. The task definition instruction of

Models	Arg	CQG	CQU	Cli	FEV	FiQA	Hot	SCI	TRE	Tou	Avg
<i>Training with Only Public data</i>											
GRITLM (Muennighoff et al., 2024)	63.17	63.68	48.54	23.24	72.57	59.91	73.96	24.41	74.31	45.66	54.94
SFR-Embedding-Mistral (Meng et al., 2024)	67.17	64.88	49.44	28.29	86.63	60.40	73.76	19.91	87.59	55.18	59.32
NV-embed-v1 (Lee et al., 2025)	68.21	66.53	52.19	27.40	84.10	63.10	78.31	20.19	85.88	55.41	60.13
LLM2Vec-Mistral-supervised (BehnamGhader et al., 2024)	57.48	66.65	49.74	21.89	70.73	53.11	54.61	22.50	77.69	38.32	51.27
GeoEmbedding (Huang et al., 2025)	77.87	64.73	50.37	42.98	92.67	53.24	73.41	22.14	77.26	54.38	60.90
BGE-en-ICL (zero-shot) (Li et al., 2025)	81.27	62.51	47.91	41.43	91.56	54.45	83.34	23.47	68.40	61.65	61.59
<i>MaxSim_{Mistral}</i>	76.91	64.45	47.46	38.64	91.14	58.20	82.94	26.30	72.36	62.10	62.04
PRISM	77.64	64.50	50.75	41.00	91.89	57.48	82.08	26.54	76.56	61.41	62.98
<i>Training with Private Data</i>											
jina-embeddings-v3 (Sturua et al., 2024)	43.29	58.02	43.52	43.14	89.90	47.35	64.70	19.92	77.74	55.28	54.28
e5-mistral(Wang et al., 2024)	61.65	63.52	46.75	28.50	86.99	56.81	73.21	16.31	87.03	55.44	57.62
stella_en_1.5B_v5 (Zhang et al., 2025a)	57.06	53.59	29.03	28.46	78.37	60.48	70.18	26.77	83.84	36.40	52.41
NV-embed-v2 (Lee et al., 2025)	70.07	65.14	51.66	33.44	90.45	65.73	83.83	21.90	88.85	57.31	62.83
gte-Qwen2-7B-instruct (Li et al., 2023a)	54.57	59.58	43.85	32.20	90.63	62.03	75.00	23.48	80.37	59.22	58.09
Linq-Embed-Mistral (Choi et al., 2024)	69.65	64.36	49.39	30.32	87.60	61.21	75.41	21.93	87.10	54.47	60.14

Table 1: Representative models on the MMTEB leaderboard, 2025. The specific meanings of the task abbreviations are shown in Appendix B.3.

Model	Latency (ms)
Vanilla	1.91
PRISM	4.12
<i>MaxSim_{Mistral}</i>	329.84

Table 2: Scoring latency over 100K candidate documents per query on a single A100 GPU.

all evaluation benchmarks can be found in Appendix B.1. For the evaluation baselines, we compare against a broad set of representative embedding models trained on similar scales of publicly available data or fine-tuned on comparable backbone LLMs. This ensures fairness by controlling for model size and data exposure. We also implement typical multi-vector retrieval strategy proposed by ColBERT (Khattab and Zaharia, 2020) on the same backbone model as baseline. The specific baselines are shown in Appendix B.2.

4.2 Main Results

Table 1 demonstrates the experiment results on retrieval subtask of MMTEB (eng, v2). We observe that our proposed **PRISM** consistently outperforms all baseline training with only public data, even achieve best among baselines training with private data. The improvements over recent LLM-based embedding approaches confirm the effectiveness of modeling documents with multi-perspective semantic representations. Specifically, compared with BGE-en-ICL (zero-shot) which is trained with the same data and setting, our model has higher retrieval performance.

Furthermore, to rigorously evaluate against exist-

ing multi-vector paradigms, we compare a strong ColBERT-style baseline *MaxSim_{Mistral}*, where implementation detail is shown in Appendix B.2. As shown in Table 1, **PRISM** consistently outperforms this typical multi-vector retrieval method. Crucially, while **PRISM** represents a document with a small set of perspective vectors which incurs a modest storage increase over single-vector baselines, it achieves a massive semantic compression compared to typical multi-vector retrieval *MaxSim_{Mistral}* that conducts late interaction among hundreds of token embeddings per query and document. This demonstrates that explicitly reasoning and extracting high-level semantic perspectives is not only more effective than exhaustive token matching, but also establishes a highly practical trade-off, where it preserves the rich expressiveness of multi-vector retrieval while drastically reducing the computational and storage overhead inherent to token-level systems. We further evaluate the inference latency over 100K candidate documents per query on a single A100 GPU. As shown in Table 2, while **PRISM** slightly increases computation over a single dot-product, its latency is comparable to the single vector baseline (Vanilla) and is significantly faster than *MaxSim_{Mistral}*.

Additionally, we evaluate **PRISM** as a general-purpose embedding model on all MMTEB tasks. The experiment results are presented in the Appendix C. It demonstrates that **PRISM** achieves higher average performance. This indicates **PRISM** can also transfer well to a broad range of semantic understanding tasks.

Models	Arg	CQG	CQU	Cli	FEV	FiQA	Hot	SCI	TRE	Tou	Avg
PRISM_{n=4}	77.64	64.50	50.75	41.00	91.89	57.48	82.08	26.54	76.56	61.41	62.98
Vanilla	81.31	62.69	47.41	44.55	92.05	54.94	81.98	28.01	67.00	59.94	61.98
w/o RE	78.84	66.00	50.98	40.38	91.52	57.22	80.44	26.41	73.54	60.44	62.57
PRISM_{n=2}	79.50	62.80	50.16	40.02	91.69	56.34	80.51	27.01	72.42	62.66	62.31
PRISM_{n=6}	79.67	62.52	49.53	43.66	90.18	57.18	81.52	26.07	58.39	63.05	61.17
PRISM_{n=8}	79.61	58.61	46.51	40.93	91.10	57.61	80.40	25.93	67.53	58.63	60.68
VIE ₁	79.33	62.01	49.21	42.90	91.89	55.10	81.24	26.81	68.78	57.74	61.50
VIE ₂	76.34	46.20	34.46	37.69	74.48	43.09	60.15	4.14	34.68	20.76	43.19
VIE ₃	76.82	40.41	31.10	38.49	82.60	34.18	68.22	1.39	27.99	7.32	40.85
VIE ₄	78.07	64.55	49.55	41.83	92.68	56.43	77.19	25.37	68.81	63.97	61.84

Table 3: Analysis of multi-perspective semantic representation. We also investigate the impact of different numbers n of semantic perspectives on retrieval performance ($n = 2, 4, 6, 8$). Further, we explore the quality for each individual semantic view (VIE_{1,2,3,4})

Models	Arg	CQG	CQU	Cli	FEV	FiQA	Hot	SCI	TRE	Tou	Avg
PRISM	77.64	64.50	50.75	41.00	91.89	57.48	82.08	26.54	76.56	61.41	62.98
LT	69.89	46.80	37.82	41.35	89.82	52.12	74.34	21.82	45.96	55.39	53.53
MP	80.24	58.82	46.99	40.06	91.64	57.88	81.66	28.92	69.71	60.94	61.68
MX	82.15	60.13	47.75	41.38	90.86	58.89	82.96	28.01	53.79	48.82	59.47

Table 4: Analysis of query-aware fusion strategy. We explore various fusion strategies, including ‘‘Last Token’’ (LT), ‘‘Mean Pooling’’ (MP) and ‘‘Maximum Value View’’ (MX).

4.3 Analysis

Analysis of multi-perspective semantic representation We first explore the significance of leveraging LLM to reason the multi-perspective semantics of documents. We compare **PRISM** with two variant denoted as ‘‘Vanilla’’ and ‘‘w/o RE’’. The ‘‘Vanilla’’ model follows the standard dense retrieval paradigm, without incorporating any reasoning instructions or semantic view tokens. In contrast, the ‘‘w/o RE’’ variant appends special semantic view tokens to documents to produce multi-perspective embeddings, yet omits the reasoning instructions. Experiment results are shown in the first two rows in Table 3. Compared with the ‘‘Vanilla’’ and ‘‘w/o RE’’ baseline, **PRISM** achieves consistently better performance, confirming that reasoning-based multiple perspective representations enhance retrieval quality beyond single-vector encodings.

Then, we examine the impact of varying the number of semantic perspectives n on retrieval performance, as shown in the middle lines in Table 3. The results indicate performance initially improves as n increases, since more perspectives allow the model to capture a broader range of semantic aspects. However, further increasing the number of perspectives leads to performance degradation. We attribute this to that excessive perspectives intro-

duce redundancy and noise, which dilute model’s ability to focus on the most salient information.

To further probe the quality of individual semantic views, we additionally evaluate the performance of **PRISM** when using each view embedding in isolation to represent documents based on overall 4 view tokens. The experiment results are shown in the last lines in Table 3. Interestingly, the results reveal a pronounced disparity across views: while the first and fourth views achieve competitive performance (61.50 and 61.84 average score), the second and third views perform substantially worse (43.19 and 40.85). Despite their overall lower averages, the second and third views achieve competitive results on certain datasets such as ArguAna. This observation suggests that weaker views at the aggregate level may still encode useful semantics under specific scenarios. Above all, the uneven quality of views underscores the necessity of both generating multiple perspectives and employing a robust fusion strategy, since relying on a single view can lead to performance fluctuations.

Analysis of Query-aware Fusion Next we explore the impact of the fusion strategies of multi-perspective semantic representations of documents on model performance. To this end, we design multiple alternative variants for comparison. Firstly,

Models	Arg	CQG	CQU	Cli	FEV	FiQA	Hot	SCI	TRE	Tou	Avg
PRISM $_{\alpha=0.01}$	77.64	64.50	50.75	41.00	91.89	57.48	82.08	26.54	76.56	61.41	62.98
w/o $\mathcal{J}_s$	79.14	64.03	50.19	39.38	89.71	55.20	80.07	26.70	68.49	60.14	61.30
PRISM $_{\alpha=0.001}$	79.41	64.09	49.76	41.51	91.47	57.92	81.56	27.05	75.32	60.15	62.82
PRISM $_{\alpha=0.1}$	77.95	62.71	50.04	39.44	90.91	56.66	81.67	26.44	72.44	58.24	61.65
PRISM $_{\alpha=0.2}$	79.85	62.84	50.67	40.56	91.17	56.55	81.38	26.86	66.47	60.04	61.63

Table 5: Analysis of contrastive semantic diversification strategy. We evaluate the performance of removing the contrastive semantic diversification loss $\mathcal{J}_s$ (w/o $\mathcal{J}_s$). Further we test the effect of different hyperparameters α on performance ($\alpha = 0.001, 0.01, 0.1, 0.2$).

Models	Arg	CQG	CQU	Cli	FEV	FiQA	Hot	SCI	TRE	Tou	Avg
PRISM $_{Qwen2.5-1.5B}$	65.01	61.33	46.14	35.14	83.84	48.76	75.19	23.11	82.19	54.92	57.56
Vanilla $_{Qwen2.5-1.5B}$	67.73	60.47	46.92	36.70	83.23	48.17	74.94	22.86	74.35	53.08	56.84
PRISM $_{Qwen2.5-3B}$	68.41	62.45	47.31	35.83	84.19	51.23	76.05	24.17	82.82	53.98	58.36
Vanilla $_{Qwen2.5-3B}$	69.95	60.61	47.26	39.85	83.15	50.43	76.95	24.87	70.97	54.56	57.86
PRISM $_{Qwen3-1.7B}$	65.03	58.51	46.50	35.10	88.37	48.97	76.21	22.88	75.43	59.01	57.60
Vanilla $_{Qwen3-1.7B}$	66.26	61.50	47.51	34.99	85.00	50.56	75.69	23.49	70.77	59.61	57.53
PRISM $_{Qwen3-4B}$	70.47	59.38	47.15	40.80	90.28	55.58	75.81	26.15	76.94	60.59	60.31
Vanilla $_{Qwen3-4B}$	70.84	60.83	51.23	39.27	88.48	54.87	75.04	26.13	66.68	58.86	59.23

Table 6: Analysis of generalizability across backbone models. We evaluate and compare the performance of PRISM and Vanilla, which is typical LLM-based dense retrieval with last special token pooling.

we append the special token $\langle eos \rangle$ after the document with the reasoning instruction and special view tokens, and directly use the hidden state of $\langle eos \rangle$ as the overall document embedding, which is denoted as “LT”. The second variant, denoted as “MP”, performs mean pooling over all special view token embeddings to construct a single comprehensive document representation. Then we utilize the representation embedding of “maximum value view” to represent document, denoted as “MX”.

Experiment results are shown in Table 4. We observe that **PRISM** consistently outperforms all variants. This demonstrates that a reliance on a single perspective (“LT” and “MX”) or leveraging average view simply (“MP”) fails to fully capture the nuanced semantics necessary for accurate retrieval. By contrast, our query-aware fusion strategy dynamically weights multiple perspectives, allowing the model to emphasize the views most pertinent to the query while still integrating complementary information. This flexible and adaptive mechanism proves more robust and effective.

Analysis of Contrastive Semantic diversification

Further, we examine the effectiveness of the proposed contrastive semantic diversification objective. We first remove the diversification loss $\mathcal{J}_s$, denoted as “w/o $\mathcal{J}_s$ ”. As shown in Table 5, the retrieval performance of **PRISM** drops significantly

without this term. This degradation indicates that the multi-perspective embeddings tend to collapse when only supervised by the standard InfoNCE loss, resulting in reduced coverage of diverse semantic perspectives. In other words, although the model still aligns queries and documents at a coarse level, it loses the ability to preserve fine-grained semantic distinctions across different perspectives, leading to weaker retrieval effectiveness.

We next explore how the weighting factor α of $\mathcal{J}_s$ affects the performance of **PRISM**. The results in Table 5 reveal a clear trade-off. Assigning too large weight to $\mathcal{J}_s$ overemphasizes diversification, which weakens alignment between queries and positive documents and introduces instability during optimization. Conversely, setting α too small causes the model to underutilize semantic diversity.

These findings confirm that contrastive semantic diversification plays a crucial role in enabling PRISM to exploit the full potential of multi-perspective representations. By appropriately regularizing view embeddings, $\mathcal{J}_s$ prevents semantic collapse and ensures that different perspectives occupy complementary regions of the embedding space. At the same time, appropriate weight α is necessary to balance alignment and diversification.

Analysis of Generalizability Across Backbone Models

To further validate the robustness of our

proposed approach, we investigate its effectiveness across different backbone models. In our experiments, we choose Qwen2.5-1.5B (Yang et al., 2024), Qwen2.5-3B (Yang et al., 2024), Qwen3-1.7B (Yang et al., 2025) and Qwen3-4B (Yang et al., 2025), which are various on model types and scales. By conducting experiments on these backbone models and comparing PRISM with the typical LLM-based dense retrieval paradigm, we observe that our multi-perspective framework consistently outperforms baseline on all backbone models. These results, which are detailed in the Table 6, further verify the broad effectiveness and generalizability of our proposed method.

5 Conclusion

In this paper, we propose **PRISM**, a LLM-based retriever which leverages reasoning capability of LLM to obtain multi-perspective document semantic representation. To ensure effective query document matching, we design a query-aware fusion strategy that adaptively integrates the relevant semantic views. Complementing this, we propose a contrastive semantic diversification objective to preserve diversity among perspectives and mitigate semantic collapse, thereby enhancing robustness across heterogeneous queries. Extensive experiments and analysis across various evaluation benchmarks demonstrate that **PRISM** achieves promising improvements in both effectiveness and generalization compared to strong baselines.

6 Limitations

Our experiments are conducted exclusively on English retrieval benchmarks. Extending multi-perspective retrieval to multilingual retrieval tasks is an important direction for our future work. Additionally, the current implementation limits the maximum input sequence length to 512 tokens and number of semantic view tokens to 4. While this is effective for common retrieval scenarios, encoding highly lengthy texts such as long-form legal case retrieval might require a dynamic scaling of the perspective tokens. A fixed small number of views may become an information bottleneck when abstracting extremely long documents that inherently contain a vast number of heterogeneous facets. Finally, while PRISM heavily compresses multi-vector retrieval from the token level down to a minimal set of semantic views (e.g., $n = 4$), it inherently introduces a linear storage and indexing

increase over strictly single vector dense retrievers. Future work will need to explore advanced representation compression techniques, such as dimension truncation or binary quantization, to optimize the storage footprint without compromising the orthogonal semantic capacities of the perspective embeddings.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62572034).

References

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. in* sem 2012: The first joint conference on lexical and computational semantics—volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, QC, Canada, pages 7–8.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd workshop on natural language processing for conversational AI*, pages 38–45.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.
- Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 34–37.
- Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A Hale, David Jurgens, and Mattia Samory. 2022. Semeval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106.
- Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy-yong Sohn. 2024. Linq-embed-mistral technical report. *arXiv preprint arXiv:2412.03223*.

- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2270–2282.
- DataCanary, hilfalkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora question pairs. <https://kaggle.com/competitions/quora-question-pairs>. Kaggle.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Çağatan, and 63 others. 2025. **Mmteb: Massive multilingual text embedding benchmark**. In *International Conference on Learning Representations*, volume 2025, pages 101715–101771.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3558–3567.
- Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. **COIL: Revisit exact lexical match in information retrieval with contextualized inverted list**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. **SimCSE: Simple contrastive learning of sentence embeddings**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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*.
- Fei Huang, Fan Wu, Zeqing Zhang, Qihao Wang, Long Zhang, Grant Michael Boquet, and Hongyang Chen. 2025. Geogpt. rag technical report. *arXiv preprint arXiv:2509.09686*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. **Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring**. In *International Conference on Learning Representations*.
- Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Shi Yu, Yishan Li, Zhiyuan Liu, Yu Gu, Ge Yu, and Maosong Sun. 2025. Learning more effective representations for dense retrieval through deliberate thinking before search. *arXiv preprint arXiv:2502.12974*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7b**. *Preprint*, arXiv:2310.06825.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.
- Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Ken Lang. 1995. Newsweeder: Learning to filter news. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. **NV-embed: Improved techniques for training LLMs as generalist embedding models**. In *The Thirteenth International Conference on Learning Representations*.
- Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Praateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftexhar Naim. 2024. **Gecko: Versatile text embeddings distilled from large language models**. *Preprint*, arXiv:2403.20327.
- Yibin Lei, Tao Shen, Yu Cao, and Andrew Yates. 2025. **Enhancing lexicon-based text embeddings with large language models**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18986–19001, Vienna, Austria. Association for Computational Linguistics.
- Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. **Meta-task prompting elicits embeddings from large language**

- models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10141–10157, Bangkok, Thailand. Association for Computational Linguistics.
- Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.
- Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. 2025. Making text embedders few-shot learners. In *International Conference on Learning Representations*, volume 2025, pages 38107–38124.
- Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023a. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Zhen Li, Chongyang Tao, Jiazhan Feng, Tao Shen, Dongyan Zhao, Xiubo Geng, and Daxin Jiang. 2023b. FAA: Fine-grained attention alignment for cascade document ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1688–1700, Toronto, Canada. Association for Computational Linguistics.
- Zhilin Liao, Xiangting Hou, Dongfang Lou, Yi-Ming Chen, and Ningyu Zhang. 2023. Camvr: context-adaptive multi-view representation learning for dense retrieval. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Xueqing Liu, Chi Wang, Yue Leng, and ChengXiang Zhai. 2018. Linkso: a dataset for learning to retrieve similar question answer pairs on software development forums. In *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering*, pages 2–5.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Maggie, Phil Culliton, and Wei Chen. 2020. Tweet sentiment extraction. <https://kaggle.com/competitions/tweet-sentiment-extraction>. Kaggle.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Wwv’18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.
- Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog*, 3:6.
- Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.
- Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.
- James O’Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish I would have loved this one, but I didn’t – a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.
- Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. [jina-embeddings-v3: Multilingual embeddings with task lora](#). *Preprint*, arXiv:2409.10173.
- 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.
- Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. 2024. LLMs are also effective embedding models: An in-depth overview. *arXiv preprint arXiv:2412.12591*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.
- Ruiran Yan, Zheng Liu, and Defu Lian. 2025. O1 embedder: Let retrievers think before action. *arXiv preprint arXiv:2502.07555*.
- 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 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. [COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bowen Zhang, Kehua Chang, and Chunping Li. 2024. Simple techniques for enhancing sentence embeddings in generative language models. In *International Conference on Intelligent Computing*, pages 52–64. Springer.
- Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025a. [Jasper and stella: distillation of sota embedding models](#). *Preprint*, arXiv:2412.19048.
- Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. [Multi-view document representation learning for open-domain dense retrieval](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000, Dublin, Ireland. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024a. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.
- Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024b. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2024. [PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4375–4391, Miami, Florida, USA. Association for Computational Linguistics.

A Training Data

Following BGE-en-ICL (Li et al., 2025), our training data contains various mixed tasks including retrieval, reranking, classification, clustering, and STS tasks.

Retrieval: ELI5 (Fan et al., 2019), HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), MSMARCO passage and document ranking (Nguyen et al., 2016), NQ (Karpukhin et al., 2020), NLI (Gao et al., 2021b), SQuAD (Karpukhin et al., 2020), TriviaQA (Karpukhin et al., 2020), Quora Duplicate Questions (DataCanary et al., 2017), Arguana (Wachsmuth et al., 2018), FiQA (Maia et al., 2018).

Reranking: SciDocsRR (Cohan et al., 2020), StackOverflowDupQuestions (Liu et al., 2018).

Classification: AmazonReviews-Classification (McAuley and Leskovec, 2013), AmazonCounterfactual-Classification (O’Neill et al., 2021), Banking77-Classification (Casanueva et al., 2020), Emotion-Classification (Saravia et al., 2018), TweetSentimentExtraction-Classification (Maggie et al., 2020), MTOPIntent-Classification (Li et al., 2021), IMDB-Classification (Maas et al., 2011), ToxicConversations-Classification (Enevoldsen et al., 2025).

Clustering: Arxiv/Biorxiv/Medrxiv/Reddit/StackExchange-Clustering-S2S/P2P, TwentyNewsgroups-Clustering (Lang, 1995).

STS: STS12 (Agirre et al., 2012), STS22 (Chen et al., 2022), STS-Benchmark (Cer et al., 2017).

The total training set includes 2.1 million samples and each sample follows the format of $\{query, d^+, \{d^-\}_m\}$, where m is the number of negative documents.

B Evaluation Details

B.1 Evaluation Instruction

The instruction for evaluation on retrieval subtask of MMTEB is shown in Table 7.

B.2 Baselines

The baselines we used for comparison include:

Training with Only Public data: GRITLM (Muennighoff et al., 2024), SFR-Embedding-Mistral (Meng et al., 2024), NV-embed-v1 (Lee et al., 2025), LLM2Vec-Mistral-supervised (BehnamGhader et al., 2024), Geoembedding (Huang et al., 2025), BGE-en-ICL (zero-shot) (Li et al., 2025). Additionally, to ensure a fair and thorough comparison with typical multi-vector retrieval paradigms, we introduce *MaxSim_{Mistral}*. This baseline is built upon the same Mistral backbone model and employs the identical training data and experimental setup as **PRISM**, but implements MaxSim late-interaction mechanism over token-level embeddings proposed by ColBERT (Khattab and Zaharia, 2020).

Training with Private Data: jina-embeddings-v3 (Sturua et al., 2024), E5-Mistral (Wang et al., 2024), stella_en_1.5B (Zhang et al., 2025a), NV-embed-v2 (Lee et al., 2025), gte-Qwen2-7B-instruct (Li et al., 2023a), Linq-Embed-Mistral (Choi et al., 2024).

B.3 Task Abbreviations

The tasks Abbreviations for MMTEB (eng, v2) retrieval subtask are as follows: Arg = ArguAna; CQG = CQADupstackGamingRetrieval; CQU = CQADupstackUnixRetrieval; Cli = ClimateFEVERHardNegatives; FEV = FEVERHardNegatives; FiQA = FiQA2018; Hot = HotpotQAHardNegatives; SCI = SCIDOCs; TRE = TRECCOVID; Tou = Touche2020Retrieval.v3; Avg represents the average scores across all datasets.

C Performance on MMTEB

We show the experiment results across all tasks in MMTEB, based on task type (Table 8) and each task (Table 9).

D Case Analysis

To better understand the framework of **PRISM**, we conduct a qualitative case analysis. Given the query-document pair, we utilize the last-layer attention map of the retriever to obtain the attention level of each view token to other document tokens. Figure 3 shows the 10 document tokens with the highest attention from each four views. It can be seen the tokens attended by view 4, which has the highest similarity score with the query, are more relevant to the query (e.g., the attended "Get an Education" can directly answer query). As a result, with our query-aware fusion strategy, views

Task Name	Instruction Template
ArguAna	Given a claim, find documents that refute the claim.
CQADupstackGamingRetrieval	Given a question, retrieve detailed question descriptions from Stackexchange that are duplicates to the given question.
CQADupstackUnixRetrieval	Given a question, retrieve detailed question descriptions from Stackexchange that are duplicates to the given question.
ClimateFEVER	Given a claim about climate change, retrieve documents that support or refute the claim.
FEVERHardNegatives	Given a claim, retrieve documents that support or refute the claim
FiQA2018	Given a financial question, retrieve user replies that best answer the question.
HotpotQAHardNegatives	Given a multi-hop question, retrieve documents that can help answer the question
SCIDOCS	Given a scientific paper title, retrieve paper abstracts that are cited by the given paper.
TRECCOVID	Given a query on COVID-19, retrieve documents that answer the query.
Touche2020Retrieval.v3	Given a question, retrieve detailed and persuasive arguments that answer the question
STS*	Retrieve semantically similar text.
SummEvalSummarization.v2	Given a news summary, retrieve other semantically similar summaries.
AmazonCounterfactualClassification	Classify a given Amazon customer review text as either counterfactual or not-counterfactual.
Banking77Classification	Given a online banking query, find the corresponding intents.
ImdbClassification	Classify the sentiment expressed in the given movie review text from the IMDB dataset.
MTOPDomainClassification	Classify the intent domain of the given utterance in task-oriented conversation.
MassiveIntentClassification	Given a user utterance as query, find the user intents.
MassiveScenarioClassification	Given a user utterance as query, find the user scenarios.
ToxicConversationsClassification	Classify the given comments as either toxic or not toxic.
TweetSentimentExtractionClassification	Classify the sentiment of a given tweet as either positive, negative, or neutral.
ArXivHierarchicalClusteringP2P	Identify the main and secondary category of Arxiv papers based on the titles and abstracts.
ArXivHierarchicalClusteringS2S	Identify the main and secondary category of Arxiv papers based on the titles.
BiorxivClusteringP2P.v2	Identify the main category of Biorxiv papers based on the titles and abstracts.
MedrxivClusteringP2P.v2	Identify the main category of Medrxiv papers based on the titles and abstracts.
MedrxivClusteringS2S.v2	Identify the main category of Medrxiv papers based on the titles.
StackExchangeClustering.v2	Identify the topic or theme of StackExchange posts based on the titles.
StackExchangeClusteringP2P.v2	Identify the topic or theme of StackExchange posts based on the given paragraphs.
TwentyNewsgroupsClustering.v2	Identify the topic or theme of the given news articles
SprintDuplicateQuestions	Retrieve duplicate questions from Sprint forum.
TwitterSemEval2015	Retrieve tweets that are semantically similar to the given tweet.
TwitterURLCorpus	Retrieve tweets that are semantically similar to the given tweet.
AskUbuntuDupQuestions	Retrieve duplicate questions from AskUbuntu forum.
MindSmallReranking	Retrieve relevant news articles based on user browsing history.

Table 7: Task instructions for MMTEB.

Models	Cla	Clu	PCla	Rer	Ret	STS	Sum	avg
<i>Training with Only Public data</i>								
gritlm	81.25	50.82	87.29	49.59	54.95	83.03	35.65	67.70
nv-embed-v1	84.11	49.50	87.05	49.16	60.13	82.20	31.40	68.32
LLM2Vec-Mistral-supervised	79.74	44.10	87.99	47.74	51.27	83.70	31.05	64.57
GeoEmbedding	89.67	56.50	82.51	48.32	60.91	80.60	30.43	70.21
BGE-en-ICL (zero-shot)	88.91	58.54	84.97	45.96	61.59	83.53	31.21	71.34
PRISM	88.59	58.81	86.18	48.00	62.98	82.28	35.24	71.74
<i>Training with Private Data</i>								
e5-mistral	79.85	51.44	88.42	49.78	57.62	84.32	36.57	67.97
stella_en_1.5B_v5	89.38	57.06	88.02	50.19	52.42	83.27	36.91	69.43
nv-embed-v2	87.19	47.66	88.69	49.61	62.84	83.82	35.21	69.81
gte-Qwen2-7B-instruct	88.52	58.97	85.9	50.47	58.09	82.69	35.74	70.72
Linq-Embed-Mistral	83.00	54.07	88.44	49.44	60.14	84.69	37.26	69.80

Table 8: Representative models on the MMTEB (eng, v2) leaderboard. Abbreviations: Cla = Classification; Clu = Clustering; PCla = Pair Classification; Rer = Reranking; Ret = Retrieval; STS = Semantic Textual Similarity; Sum = Summarization. Avg represents the average score across all tasks.

Dataset	E5-mistral	NV-enbed-v2	BGE-en-ICL (zero-shot)	PRISM
AmazonCounterfactualClassification	74.57	79.28	92.70	92.67
ArXivHierarchicalClusteringP2P	65.28	60.28	64.31	65.67
ArXivHierarchicalClusteringS2S	61.28	59.24	62.84	62.78
ArguAna	61.65	70.07	81.27	77.64
AskUbuntuDupQuestions	66.98	67.46	61.44	63.93
BIOSSES	85.50	87.42	87.76	86.00
Banking77Classification	81.41	92.42	90.85	90.68
BiorxivClusteringP2P.v2	42.44	43.73	54.05	53.93
CQADupstackGamingRetrieval	63.52	65.14	62.51	64.50
CQADupstackUnixRetrieval	46.75	51.66	47.91	50.75
ClimateFEVERHardNegatives	28.50	33.44	41.43	41.00
FEVERHardNegatives	86.99	90.45	91.56	91.89
FiQA2018	56.81	65.73	54.45	57.48
HotpotQAHardNegatives	73.21	83.83	83.34	82.08
ImdbClassification	94.73	97.14	96.82	96.61
MTOPDomainClassification	95.33	95.69	97.19	97.43
MassiveIntentClassification	77.00	78.28	80.32	79.66
MassiveScenarioClassification	79.13	81.12	81.81	81.84
MedrxivClusteringP2P.v2	39.05	37.40	48.65	48.24
MedrxivClusteringS2S.v2	37.82	35.51	46.58	46.79
MindSmallReranking	32.59	31.76	30.49	33.00
SCIDOCS	16.31	21.90	23.47	26.54
SICK-R	82.64	82.15	84.24	82.87
STS12	79.65	77.89	77.82	74.81
STS13	88.43	88.30	85.38	83.35
STS14	84.54	84.30	81.30	80.32
STS15	90.42	89.04	87.17	86.84
STS17	91.76	90.67	91.38	91.14
STS22.v2	67.29	66.20	69.61	68.84
STSBenchmark	88.60	88.41	87.13	86.35
SprintDuplicateQuestions	95.86	97.09	95.12	96.97
StackExchangeClustering.v2	64.33	55.04	76.68	77.44
StackExchangeClusteringP2P.v2	48.07	45.00	49.36	49.41
SummEvalSummarization.v2	36.57	35.21	31.21	35.24
TRECCOVID	87.03	88.85	68.40	76.56
Touche2020Retrieval.v3	55.44	57.31	61.65	61.41
ToxicConversationsClassification	71.75	92.74	93.09	91.05
TweetSentimentExtractionClassification	64.89	80.87	78.53	78.78
TwentyNewsgroupsClustering.v2	53.28	45.12	65.86	66.26
TwitterSemEval2015	81.62	81.11	73.96	75.29
TwitterURLCorpus	87.79	87.87	85.85	86.28
Avg	67.97	69.81	71.34	71.74

Table 9: Detail results in MMTEB (eng, v2).

that capture more effective semantics account for a larger proportion during fusion. Meanwhile, the semantics attended by different views are quite diverse, which can provide supplementary semantics in matching process.

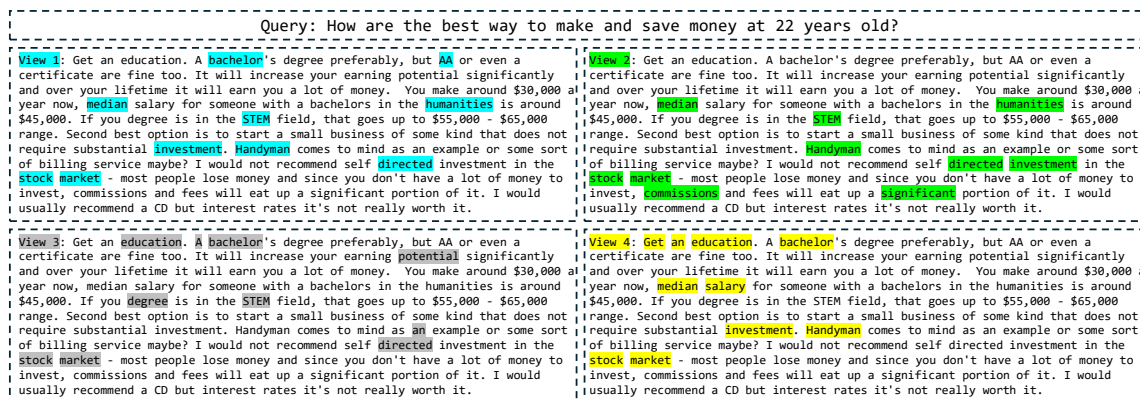


Figure 3: Given a query and a document, we show the attention level of each view token in the document to other tokens and highlight the top-10 tokens. Meanwhile, the similarity scores between each view and query embedding are calculated, with views 1 through 4 as follows: 0.2478, 0.1993, 0.1931 and 0.3598.