

AnInter: Anti-Interference Training for Long Context Retrieval

Lijie Ye

South China University of Technology
yelijie33@gmail.com

Abstract

Dense retrievers in Retrieval-Augmented Generation (RAG) have achieved substantial progress in semantic matching, but they remain limited in long context retrieval. In long context retrieval, target evidence often occupies only a small local span, while the remaining content acts as noise. We aim to improve long-document retrieval by enhancing retrievers' anti-interference capability. Our contributions are threefold: (1) we construct the Anti-Interference Capability Benchmark (AICB) to evaluate retrievers' anti-interference capability under controlled noise levels; (2) we propose anti-interference training as a short-text training path for improving long context retrieval; and (3) we train AnInter, a series of embedding models that are more robust to noisy long documents. Evaluation on LongEmbed shows that AnInter improves over its backbone by 16.4 points on average and achieves state-of-the-art performance on this challenging long context embedding benchmark. These results show that interference-aware training is a practical direction for improving long context retrieval.

1 Introduction

Retrieval-augmented generation (RAG) systems rely on retrievers to access external evidence for downstream generation (Lewis et al., 2020). As RAG systems are increasingly applied to legal documents, enterprise knowledge bases, scientific papers, technical manuals, and meeting transcripts, retrievers are required to handle long documents rather than only short text passages. Recent retrievers, such as the Qwen3-Embedding series, have extended the context window to 32K tokens (Zhang et al., 2025). However, context-window extension and contrastive training with long texts have not significantly improved long-text retrieval ability. In our experiments, Qwen3-Embedding-8B performs worse on the challenging LongEmbed benchmark

than E5-Mistral, which only has a 4096-token context length (Wang et al., 2024; Zhu et al., 2024).

Existing studies improve long-text retrieval through context-window extension (Zhu et al., 2024; Zhang et al., 2025), long-context encoder design (Saad-Falcon et al., 2024), contextual chunk representations (Günther et al., 2024), and multi-vector retrieval (Khattab and Zaharia, 2020; Santhanam et al., 2022; Faysse et al., 2024). These studies suggest that long-text retrieval requires more than short-passage matching, because dense retrievers can still fail when useful evidence is embedded in long documents. Therefore, improving the discriminability of long-text embeddings at the retriever level is an unavoidable challenge.

We observe that the difficulty of long context retrieval lies in the following fact: target evidence often occupies only a small part of a document and is surrounded by other content. Therefore, an effective long-text retriever should have the ability to distinguish the target evidence from distracting content within a long retrieval unit.

As shown in Figure 1(a), standard contrastive learning mainly learns isolated query-document matching: each query is pulled close to its matched document and pushed away from other documents. However, if a mixed document contains both d_1 and d_2 , it should still be relevant to q_1 because it contains the evidence from d_1 . Such mixed positives are not explicitly constructed in standard training data. Therefore, enabling the retriever to maintain stable vector similarity as distracting content increases is key to improving long context retrieval ability.

To address this problem, we construct anti-interference data by combining origin documents, as illustrated in Figure 1(b), and train the retriever with the expanded matching matrix shown in Figure 1(c).

Our contributions can be summarized as follows:

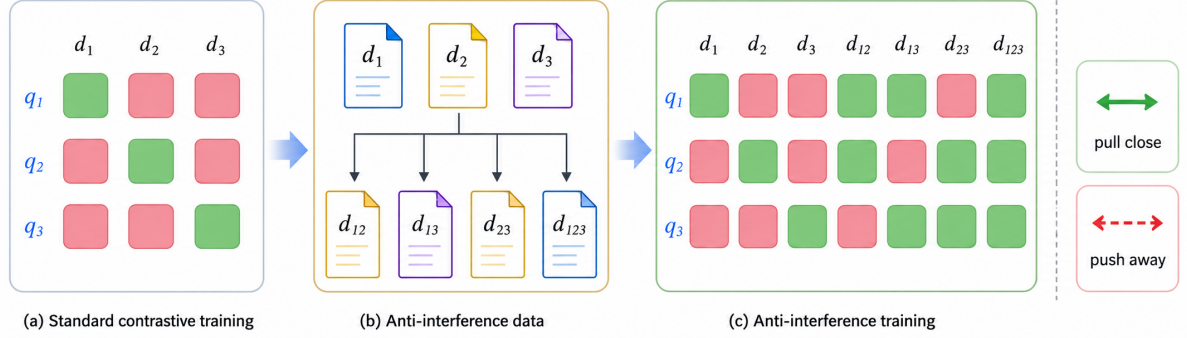


Figure 1: Illustration of standard contrastive training, anti-interference data, and anti-interference training. (a) Standard contrastive training pulls each query close to its matched document and pushes it away from other documents. (b) Anti-interference data combines origin documents into mixed documents. (c) Anti-interference training expands the training target from origin documents to mixed documents that still contain the query’s origin evidence.

- We are the first to study anti-interference retrieval for dense retrievers. We construct AICB, an Anti-Interference Capability Benchmark, to evaluate retrievers under controlled noise levels.
- We propose anti-interference training as a short-text training path to improve long context retrieval. It teaches the retriever to preserve query-document matching ability under interference.
- We develop AnInter by fine-tuning Qwen3 series models with anti-interference training. AnInter achieves a **16.4** point average gain over the backbone on LongEmbed and outperforms E5-Mistral + NTK by **6.4** points.

2 Related Work

2.1 Dense Retrieval and Contrastive Training

Dense retrieval represents queries and documents in a shared embedding space and ranks documents by vector similarity. DPR (Karpukhin et al., 2020) trains dual encoders for open-domain question answering. Contriever (Izacard et al., 2022) extends dense retrieval to unsupervised contrastive learning. E5 (Wang et al., 2022) uses weakly supervised contrastive pre-training over large-scale text pairs. GTR (Ni et al., 2022) studies scaled dual encoders with a fixed-size dot-product bottleneck. BGE/C-Pack (Xiao et al., 2024) further provides large-scale embedding resources based on relevance-oriented training signals.

Standard contrastive learning usually trains one query with one positive passage and in-batch negatives (van den Oord et al., 2018; Karpukhin et al., 2020). This setting does not model the noise caused by real document segmentation, where target evidence can be mixed with other content in the same document. Long-text contrastive training is also expensive because mainstream dense retrievers are Transformer-based (Vaswani et al., 2017; Devlin et al., 2019; Karpukhin et al., 2020), and self-attention has quadratic cost with sequence length (Tay et al., 2022). Longer inputs reduce the feasible batch size and weaken in-batch negative learning. Our work therefore uses short-text anti-interference training to improve long-text retrieval ability more efficiently.

2.2 Long-Text Retrieval

LongEmbed (Zhu et al., 2024) extends embedding models to longer context windows. LoCo/M2-BERT (Saad-Falcon et al., 2024) designs long-context retrieval encoders for 32K-token documents. Late Chunking (Günther et al., 2024) first encodes the full document and then obtains contextual chunk embeddings. ColBERT (Khattab and Zaharia, 2020), ColBERTv2 (Santhanam et al., 2022), and ColPali (Faysse et al., 2024) use multi-vector representations to preserve fine-grained matching signals.

These methods improve context length, chunk representation, or matching granularity. Qwen3-Embedding (Zhang et al., 2025) further supports long-context embedding through multi-stage training. In our evaluation, Qwen3-Embedding still

performs poorly on LongEmbed despite its long-context support. Therefore, improving the discriminability of long-text embeddings at the retriever level remains an unavoidable challenge. We address this problem from the training perspective. By injecting anti-interference capability during training, we improve its performance on long-text retrieval tasks.

3 Anti-Interference Capability Benchmark

In this section, we first introduce an automatic method for building the Anti-Interference Capability Benchmark (AICB), including the collection and filtering of origin query-document pairs and the synthesis of anti-interference data. We refer to the synthesized mixed documents as noisy positive documents. We then introduce our metric system for quantifying the anti-interference capability of models.

3.1 Collection of Origin Dataset

We sample 25,000 candidate pairs from MS MARCO v1.1 (Bajaj et al., 2016) and SQuAD (Rajpurkar et al., 2016). The answers in these candidate query-document pairs are all 100–250 words long. We then conduct two filtering steps. First, we select six representative embedding models from the MTEB leaderboard (Muennighoff et al., 2023) whose scores fall in the range of 50–80: Qwen3-Embedding-0.6B/4B/8B (Zhang et al., 2025), LLaMA-Embed-Nemotron-8B (Babakhin et al., 2025), Instructor-XL (Su et al., 2023) and NV-Embed-v2 (Lee et al., 2024). We evaluate these six models on the 25,000 query-document pairs and retain the pairs that most models can retrieve correctly. Finally, we retain 11,450 pairs. On these 11,450 query-document pairs, all six models achieve Hit@1 above 80%. This step ensures that the models already have the ability to distinguish these query-document pairs, so that the later performance drop mainly reflects the interference effect, rather than weak origin matching ability.

Second, to ensure that documents can serve as interference documents for each other, we compute the cosine similarity gap between the gold document and the second-ranked candidate document, and only retain pairs whose minimum gap across all models is larger than 0.15. The final origin dataset contains 5,223 query-document pairs. We call the original matched document the gold

document. In this way, we try to ensure that the other documents can act as interference documents for the gold document, rather than being related documents.

3.2 Synthesis of Noisy Positive Documents

As shown in Algorithm 1, we use the original query-document pair to synthesize noisy positive documents. First, we introduce noise level k to quantify the amount of interference. For example, $k = 3$ means that three noisy documents are added to the gold document. For each noise level k , we apply a circular adjacent window of size $k + 1$ to concatenate the gold document with k neighboring origin documents. The generated document is called a noisy positive document. The circular adjacent window is used to ensure that different queries have the same number of documents at different noise levels and avoid duplicate documents. Notably, we randomize the document order when generating noisy positive documents.

In the design of AICB, the 5,223 origin query-document pairs are split into 3,656 training queries, 522 validation queries, and 1,045 test queries. We use thirteen noise levels, $k = \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60\}$, to build the test dataset. For the training and validation datasets, we use five noise levels, $k = \{1, 2, 3, 4, 5\}$, following Algorithm 1.

3.3 Metric

Measuring a model’s anti-interference capability requires two conditions. First, the model should be able to correctly match the origin query-document pair. Second, after interference documents are added, the noisy positive documents should still rank above all negative documents. Following these two ideas, we design the metric TanIn.

For query q_i , let d_i denote the origin positive document, and let $s(q, d)$ denote the similarity score produced by the retriever. We define the negative document set as $\mathcal{N}_i$ and the strongest negative baseline as

$$b_i = \max_{d \in \mathcal{N}_i} s(q_i, d). \quad (1)$$

The origin matching indicator is

$$c_i = \mathbf{1}[s(q_i, d_i) > b_i]. \quad (2)$$

Only queries with $c_i = 1$ are used for anti-interference evaluation. We compute Ori.Hit@1 as the average origin matching accuracy over all

Algorithm 1: Synthesis of Noisy Positive Documents

Input: Origin query-document pairs

$\mathcal{B} = \{(q_i, d_i)\}_{i=1}^N$; noise levels $\mathcal{K}$;
random seed s .

Output: Corpus $\mathcal{D}_{\text{all}}$ and qrels $\mathcal{Q}_{\text{rel}}$.

```
1  $\mathcal{D}_0 \leftarrow \text{Unique}(\{d_i\}_{i=1}^N)$ ;  
2  $\mathcal{D}_{\text{all}} \leftarrow \mathcal{D}_0$ ;  
3  $\mathcal{Q}_{\text{rel}} \leftarrow \{(q_i, d_i)\}_{i=1}^N$ ;  
4  $\pi \leftarrow \text{Shuffle}(\mathcal{D}_0; s)$ ;  
5  $\pi_t$  denotes the  $t$ -th document in the shuffled  
   corpus;  
6 foreach  $k \in \mathcal{K}$  do  
7   for  $t \leftarrow 1$  to  $|\mathcal{D}_0|$  do  
8      $x_{k,t} \leftarrow \text{Concat}(\pi_t, \dots, \pi_{t+k})$ ;  
     // circular indexing  
9     Add  $x_{k,t}$  to  $\mathcal{D}_{\text{all}}$  if it has not  
     appeared before;  
10    foreach  $(q_i, d_i) \in \mathcal{B}$  such that  
11       $d_i \in \{\pi_t, \dots, \pi_{t+k}\}$  do  
        Add  $(q_i, x_{k,t})$  to  $\mathcal{Q}_{\text{rel}}$  as a  
        positive label;  
12 return  $\mathcal{D}_{\text{all}}, \mathcal{Q}_{\text{rel}}$ ;
```

queries:

$$\text{Ori.Hit@1} = \frac{1}{N} \sum_{i=1}^N c_i. \quad (3)$$

Let $\mathcal{I} = \{i \mid c_i = 1\}$ denote these valid queries. For each noise level k , let $\mathcal{P}_{i,k}^{\text{noise}}$ denote the noisy positive set of q_i . The query-level score is

$$r_i^{(k)} = \frac{1}{|\mathcal{P}_{i,k}^{\text{noise}}|} \sum_{d \in \mathcal{P}_{i,k}^{\text{noise}}} \mathbf{1}[s(q_i, d) > b_i]. \quad (4)$$

The anti-interference score at noise level k is

$$\text{anIn}@k = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} r_i^{(k)}. \quad (5)$$

The overall TanIn score is averaged over all noise levels:

$$\text{TanIn} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \text{anIn}@k. \quad (6)$$

4 Anti-Interference Training

We use Multi-Positive InfoNCE for anti-interference training. In standard contrastive

training, each query usually has only one positive document. In our setting, however, a query should match not only its origin evidence document, but also every noisy positive document. Therefore, for each query q_i , we define a positive set $\mathcal{P}_i$ consisting of the original document and all its noisy positive documents. All candidate documents outside $\mathcal{P}_i$ are treated as negatives.

Training is performed at the level of each positive document. For every $p \in \mathcal{P}_i$, we form a contrastive objective between p and the in-batch negatives $\mathcal{N}_i$:

$$\ell(q_i, p) = -\log \frac{r(q_i, p)}{r(q_i, p) + \sum_{n \in \mathcal{N}_i} r(q_i, n)}. \quad (7)$$

The final Multi-Positive InfoNCE loss averages this positive-level loss over all queries and all their positives:

$$\mathcal{L}_{\text{MPI}} = \text{Avg}_{i, p \in \mathcal{P}_i} \ell(q_i, p). \quad (8)$$

Here, $r(q, d) = \exp(s(q, d)/\tau)$, $s(q, d)$ is the cosine similarity between normalized embeddings, and τ is the temperature.

5 Experiments

5.1 Experiment Setup

Training AnInter. All experiments are conducted on 8 RTX PRO 6000 GPUs. We use the Qwen3-Embedding series (Zhang et al., 2025) as the backbone retrievers. For training data construction, we use noise levels $k = \{1, 2, 3, 4, 5\}$. Under the circular window construction, each origin document appears in $k + 1$ noisy positives at noise level k . Therefore, using $k = \{1, 2, 3, 4, 5\}$ gives $2 + 3 + 4 + 5 + 6 = 20$ noisy positives for each query. At each training step, each GPU is assigned one query, its 21 positive documents, and several negative documents. To avoid out-of-memory errors, we sample 32, 8, and 4 negative documents for Qwen3-Embedding-0.6B, Qwen3-Embedding-4B, and Qwen3-Embedding-8B, respectively. We then gather document embeddings across all GPUs to construct the global candidate pool for the current training step. The multi-positive loss is calculated over this global candidate pool. We apply LoRA (Hu et al., 2022) with ranks 128, 256, and 512 for Qwen3-Embedding-0.6B, 4B, and 8B, respectively, yielding AnInter-0.6B, AnInter-4B, and AnInter-8B. Detailed training hyperparameters are provided in Appendix Table 7.

Model	Ori.Hit@1	anIn@1	anIn@5	anIn@10	anIn@20	anIn@30	anIn@40	anIn@50	anIn@60	TanIn
Baseline										
llama-embed-nemotron-8B	83.6	50.3	8.2	4.1	1.8	1.2	0.8	0.5	0.4	5.6
NV-Embed-v2	95.4	68.8	14.1	4.7	1.7	0.9	0.3	0.3	0.9	7.4
E5-Mistral-7B-instruct	71.0	33.5	10.9	5.9	3.1	2.0	1.5	1.2	1.0	5.4
GTE-Qwen2-1.5B-instruct	53.7	23.3	7.7	4.2	2.2	1.5	1.1	0.9	0.8	3.8
GritLM-7B	52.6	21.9	5.8	3.0	1.6	1.1	0.8	0.7	0.6	3.2
instructor-xl	99.0	77.7	<u>29.6</u>	<u>16.1</u>	<u>8.4</u>	<u>5.7</u>	<u>4.3</u>	<u>3.5</u>	<u>2.9</u>	<u>13.7</u>
Qwen3-Embedding-0.6B	95.9	58.9	10.3	5.0	2.2	2.5	3.1	2.6	2.5	7.7
Qwen3-Embedding-4B	98.9	77.7	21.2	11.2	5.2	3.5	2.8	1.8	1.3	11.0
Qwen3-Embedding-8B	<u>99.3</u>	<u>79.1</u>	24.6	8.0	3.2	2.2	1.7	1.5	1.4	10.3
Our AnInter										
AnInter-0.6B	90.0	78.6	53.6	44.3	28.0	21.2	17.9	15.2	13.3	29.3
AnInter-4B	95.2	90.2	74.8	64.9	46.6	36.8	30.4	26.4	23.7	44.3
AnInter-8B	96.8	87.7	80.4	70.6	47.7	37.5	33.4	30.2	27.1	46.9
Δ -Best baseline	-2.5	+11.1	+50.8	+54.5	+39.3	+31.8	+29.1	+26.7	+24.2	+33.2

Table 1: **Main results on AICB.** We compare AnInter with strong embedding baselines on AICB. Ori.Hit@1 measures retrieval accuracy on origin query-document pairs, while anIn@k and TanIn measure robustness under increasing interference levels. Bold numbers indicate the best result in AnInter models and underlined numbers indicate the best result in nine baselines. For readability, we report Ori.Hit@1, TanIn, and representative anIn@k scores in the main table, including anIn@1, anIn@5, anIn@10, anIn@20, anIn@30, anIn@40, anIn@50, and anIn@60. The complete results over all thirteen noise levels are provided in Appendix Table 4.

Evaluation on AICB. We evaluate AnInter on AICB to verify its anti-interference retrieval performance. In addition to the Qwen3-Embedding series backbones, we select LLaMA-Embed-Nemotron-8B (Babakhin et al., 2025), NV-Embed-v2 (Lee et al., 2024), E5-Mistral-7B-instruct (Wang et al., 2024), GTE-Qwen2-1.5B-instruct (Li et al., 2023), GritLM-7B (Muennighoff et al., 2024), and Instructor-XL (Su et al., 2023) as baselines. These models obtain MTEB Multilingual scores (Enevoldsen et al., 2025) between 50 and 80, representing different levels of semantic matching ability.

Evaluation on LongEmbed. LongEmbed (Zhu et al., 2024) is a challenging benchmark for long-context document retrieval. We evaluate AnInter on LongEmbed to examine the relationship between anti-interference capability and long-document retrieval ability. We include representative baselines from the original LongEmbed evaluation, including 512-context embedding models such as E5Base (Wang et al., 2022), GTEBase (Li et al., 2023), and Contriever (Izacard et al., 2022), as well as long-context embedding models such as E5-Mistral (Wang et al., 2024), Jina-V2 (Günther et al., 2023), Nomic-V1 (Nussbaum et al., 2024), and BGE-M3 (Chen et al., 2024). We also include context-extension baselines, including E5Base + Tuning (Wang et al., 2022), E5-RoPEBase + SelfExtend (Jin et al., 2024), and E5-Mistral + NTK.

Evaluation on standard retrieval benchmarks.

To examine the impact of anti-interference training on general semantic matching ability, we evaluate the models on ten retrieval benchmarks included in MTEB (Muennighoff et al., 2023). The tasks include ArguAna (Arg) (Wachsmuth et al., 2018), CQADupStack-Gaming (CQG) (Hoogeveen et al., 2015), CQADupStack-Unix (CQU), ClimateFEVER-HN (CFHN) (Diggelmann et al., 2020; Enevoldsen et al., 2025), FEVER-HN (FEV) (Thorne et al., 2018), FiQA2018 (FiQA) (Maia et al., 2018), HotpotQA-HN (HPQA) (Yang et al., 2018), SCIDOCS (SCID) (Cohan et al., 2020), TRECCOVID (TREC) (Voorhees et al., 2021; Wang et al., 2020), and Touché 2020 (TOU) (Bondarenko et al., 2020).

5.2 Main Results

Results on AICB. We first evaluate all models on AICB, with the results summarized in Table 1. The baseline models achieve high Ori.Hit@1 scores, but their performance drops sharply when noisy positives are introduced. Qwen3-Embedding-8B reaches 99.3 Ori.Hit@1, but its anIn@5 drops to 24.6. This shows that strong origin semantic matching does not directly imply robustness under interference. AnInter-8B achieves the best TanIn score of 46.9, improving the best baseline by 33.2 points. These results demonstrate that AnInter substantially improves robustness against noisy positive documents while keeping Ori.Hit@1 at a high level.

Model	Syn.		Real				Avg.
	Needle	Passkey	QMS	SFD	WQA	NQA	
Baseline							
E5Base	28.5	38.0	23.8	74.7	55.8	25.3	41.0
GTEBase	24.5	31.0	21.8	55.8	47.3	28.6	34.8
Contriever	29.0	38.5	25.5	73.5	47.3	26.7	40.1
E5-Mistral	48.3	71.0	43.6	96.8	82.0	44.6	64.4
Jina-V2	54.5	50.3	38.9	93.5	74.0	37.9	58.2
Nomic-V1	39.5	60.7	36.7	93.0	73.8	41.2	57.5
BGE-M3	40.5	59.3	35.5	94.0	78.0	45.8	58.9
Qwen3-0.6B	38.5	92.8	40.1	92.2	81.2	24.2	61.5
Qwen3-4B	59.0	98.5	40.8	94.9	84.9	26.8	67.5
Qwen3-8B	48.5	93.3	39.8	82.7	57.6	19.1	56.8
E5Base + Tuning	41.5	67.3	35.7	95.2	69.2	30.4	56.6
E5-RoPE + SE	53.5	73.5	39.1	91.9	74.6	32.3	60.8
E5-Mistral + NTK	<u>66.8</u>	93.8	<u>49.2</u>	<u>97.1</u>	<u>95.2</u>	<u>49.8</u>	<u>75.3</u>
Our AnInter							
AnInter-0.6B	74.8	98.5	44.6	95.4	90.3	46.3	75.0
AnInter-8B	86.0	99.8	55.5	94.6	79.4	54.0	78.2
AnInter-4B	91.8	98.0	49.6	98.1	91.1	61.3	81.7
Δ -Backbone	+35.5	+3.9	+9.7	+6.1	+12.4	+30.5	+16.4
Δ -Best	+25.0	+1.3	+6.3	+1.0	-4.1	+11.5	+6.4

Table 2: **Results on LongEmbed.** We evaluate AnInter on the LongEmbed benchmark, which includes two synthetic retrieval tasks and four real long-context retrieval tasks. QMS, SFD, WQA, and NQA denote QMSum, SummScreenFD, 2WikiMultihopQA, and NarrativeQA, respectively. The synthetic tasks are reported with Acc@1, while the real tasks are reported with nDCG@10. We include representative LongEmbed baselines, including 512-context embedding models, long-context embedding models, and context-extension baselines. Bold numbers indicate the best result in AnInter models and underlined numbers indicate the best result among all compared baselines.

Results on LongEmbed. We next evaluate the transferability of anti-interference capability to LongEmbed. The results are summarized in Table 2. On average, AnInter improves the corresponding backbones by 16.4 points. AnInter-4B achieves the best overall score of 81.7, outperforming the strongest LongEmbed baseline, E5-Mistral + NTK, by 6.4 points. The gains are especially large on Needle and NarrativeQA, where AnInter improves over the best baseline by 25.0 and 11.5 points, respectively. These results suggest that anti-interference training is not limited to AICB. It also improves the model’s ability to locate useful evidence from long documents.

Results on retrieval benchmarks. We further evaluate whether AnInter preserves general retrieval ability, with results shown in Table 3. Compared with the corresponding Qwen3-Embedding backbones, AnInter-0.6B, AnInter-4B, and AnInter-8B reduce the average score by 6.9, 3.1, and 1.5 points, respectively. This moderate drop is expected be-

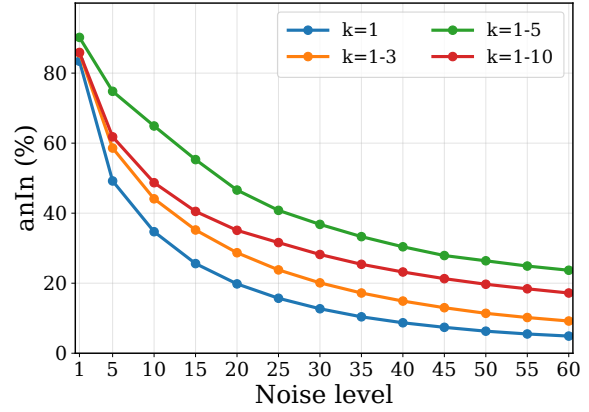


Figure 2: **Effect of training noise levels.** We compare different choices of training noise levels on AICB. Qwen3-Embedding-4B fine-tuned with moderate noise levels, especially $k = \{1, 2, 3, 4, 5\}$, achieves stronger robustness across test interference levels than training with only weak noise or using a broader noise range $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

cause AnInter changes conventional one-to-one contrastive learning (van den Oord et al., 2018; Karpukhin et al., 2020) into a multi-positive objective, which can disperse the original semantic matching signal. However, this cost is small compared with the gains on AICB and LongEmbed. AnInter-8B still achieves a competitive average score of 67.9, ranking first among all non-backbone models. This shows that AnInter improves interference-aware retrieval while largely preserving general semantic matching ability.

6 Ablation Studies

Effect of training noise levels. We first study how training noise levels affect anti-interference learning. In this experiment, we fine-tune Qwen3-Embedding-4B under different training noise settings, including weak interference $k = \{1\}$, moderate interference $k = \{1, 2, 3\}$ and $k = \{1, 2, 3, 4, 5\}$, and a broader setting $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The results are shown in Figure 2, and the full scores are reported in Appendix Table 5. Training with $k = \{1, 2, 3, 4, 5\}$ achieves the best robustness among the single-stage settings across test interference levels. In contrast, $k = \{1\}$ mainly improves performance under weak interference, while the broader setting $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ makes optimization less stable and reduces Ori.Hit@1. This shows that moderate and focused noise levels are more effective for anti-interference training.

Model	Arg	CQG	CQU	CFHN	FEV	FiQA	HPQA	SCID	TREC	TOU	Avg.
NV-Embed-v2	70.1	65.1	51.7	33.4	90.4	65.7	83.8	21.9	88.9	57.3	62.8
E5-Mistral-7B-instruct	61.7	63.5	46.8	28.5	87.0	56.8	73.2	16.3	87.0	55.4	57.6
GTE-Qwen2-1.5B-instruct	69.7	61.3	45.0	13.2	27.9	54.7	70.4	25.0	85.4	50.0	50.3
GritLM-7B	63.2	63.7	48.5	23.2	72.6	59.9	74.0	24.4	74.3	45.7	54.9
Qwen3-Embedding-0.6B	71.0	64.1	51.5	43.6	88.9	46.6	67.7	24.4	90.5	69.9	61.8
Qwen3-Embedding-4B	<u>75.6</u>	71.5	<u>59.6</u>	48.5	<u>92.5</u>	62.7	75.2	<u>31.4</u>	<u>92.9</u>	74.6	<u>68.5</u>
Qwen3-Embedding-8B	76.9	<u>72.4</u>	61.2	49.0	92.7	<u>64.6</u>	77.0	32.7	95.0	<u>73.0</u>	69.4
AnInter-0.6B	67.3	64.3	46.3	44.1	81.7	33.5	70.1	15.8	73.8	52.3	54.9
AnInter-4B	74.9	72.3	58.2	<u>49.7</u>	90.1	56.8	78.3	27.9	79.2	66.9	65.4
AnInter-8B	74.4	74.9	58.6	53.4	89.9	59.4	<u>81.3</u>	<u>31.4</u>	86.8	68.8	67.9

Table 3: **Results on ten retrieval tasks.** We evaluate AnInter on ten retrieval benchmarks included in MTEB to examine whether anti-interference fine-tuning preserves general retrieval ability. The last column reports the average score across all tasks. The tasks include ArguAna (Arg), CQADupStack-Gaming (CQG), CQADupStack-Unix (CQU), ClimateFEVER-HN (CFHN), FEVER-HN (FEV), FiQA2018 (FiQA), HotpotQA-HN (HPQA), SCIDOCs (SCID), TRECCOVID (TREC), and Touché 2020 (TOU). Bold numbers indicate the best result and underlined numbers indicate the second-best result.

Effect of origin data anchoring. We then examine whether origin query-document pairs are necessary for stable anti-interference learning. Starting from the standard $k = \{1, 2, 3, 4, 5\}$ training setting on Qwen3-Embedding-4B, we compare it with a variant that removes origin query-document pairs and trains only with noisy positive documents. The results are shown in Figure 3, and the full scores are reported in Appendix Table 5. Removing origin data reduces TanIn from 44.3 to 32.5 and also lowers Ori.Hit@1 from 95.2 to 91.9. This indicates that origin documents serve as semantic anchors. They help the model preserve the correct query-document relation while learning from noisy positives.

Effect of backbone origin semantic matching ability. Next, we study whether anti-interference learning depends on the backbone’s origin retrieval ability. For this experiment, we use Qwen3-Embedding-4B to construct training subsets with different backbone Ori.Hit@1 values, while keeping the number of training queries the same for fairness. Each subset is then used to train an anti-interference model with $k = \{1, 2, 3, 4, 5\}$ and the same LoRA configuration as AnInter-4B. The results are shown in Figure 4, and the full scores are reported in Appendix Table 6. As the backbone Ori.Hit@1 of the training subset increases from 37.6 to 98.9, the final TanIn after fine-tuning increases from 8.4 to 44.3. This indicates that anti-interference capability is built upon the backbone’s origin semantic matching ability.

Effect of progressive multi-stage training. We further examine whether progressive training can

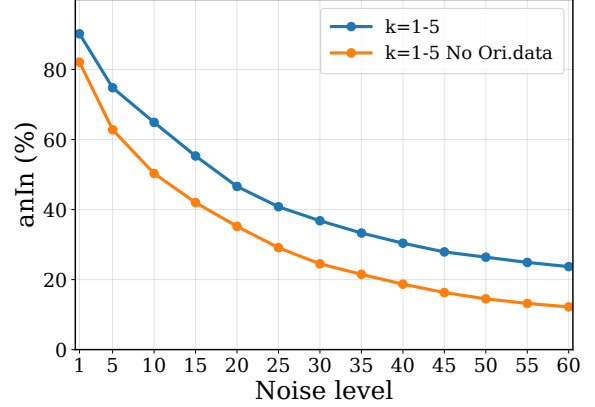


Figure 3: **Effect of origin data anchoring.** We compare training with and without origin data under the $k = \{1, 2, 3, 4, 5\}$ setting. Removing origin data weakens Qwen3-Embedding-4B performance, showing that origin query-document pairs provide an anchoring effect for stable anti-interference learning.

improve robustness. In this experiment, we compare direct training on the full noise range $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ with a progressive strategy that first trains Qwen3-Embedding-4B with $k = \{1, 2, 3, 4, 5\}$ and then continues training with $k = \{6, 7, 8, 9, 10\}$. The results are shown in Figure 5, and the full scores are reported in Appendix Table 5. Progressive training improves TanIn from 35.2 to 47.4, suggesting that AnInter can first learn stable matching under moderate interference and then adapt to harder noisy positives. This shows that anti-interference capability can be strengthened step by step. Since higher noise levels make noisy positives longer and slow both forward and backward passes, practical training should balance

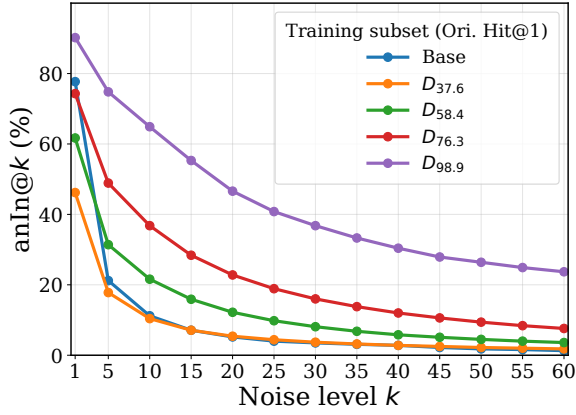


Figure 4: **Effect of backbone origin retrieval ability.** We construct training subsets with different backbone Ori.Hit@1 values using Qwen3-Embedding-4B and compare their anti-interference performance after fine-tuning with $k = \{1, 2, 3, 4, 5\}$. The results show that stronger origin semantic matching ability leads to better anti-interference robustness.

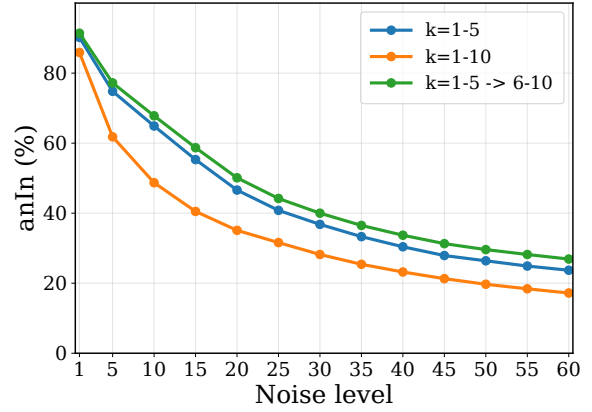


Figure 5: **Effect of progressive multi-stage training.** We compare direct training on the full noise range $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ with progressive training from $k = \{1, 2, 3, 4, 5\}$ to $k = \{6, 7, 8, 9, 10\}$. Progressive training achieves better robustness across interference levels, suggesting that gradually increasing interference difficulty is more effective than exposing the model to all noise levels at once.

robustness gains and computational cost.

7 Conclusion

This paper explores anti-interference training for dense retrievers. We show that high origin retrieval accuracy does not guarantee robustness when target evidence is mixed with interfering content, revealing a limitation of conventional semantic matching. To address this issue, we train AnInter with both origin documents and noisy positives under a multi-positive objective. Experiments show that AnInter substantially improves anti-interference robustness and transfers this ability to long-context retrieval on LongEmbed. Further analysis shows that effective anti-interference training depends on origin-data anchoring, suitable noise levels, and the backbone’s original semantic matching ability. The effectiveness of progressive training further suggests that AnInter still has room for improvement. These findings indicate that future retriever training should explicitly model noisy positive interference, especially for real-world RAG scenarios (Lewis et al., 2020) where evidence is often long, mixed, and noisy.

Limitations

As an initial study of anti-interference training for dense retrievers, this work is limited in several aspects. First, AICB constructs noisy positive documents by combining origin query-document pairs under controlled noise levels. This design

enables systematic evaluation of anti-interference capability, but it may not fully cover the diversity of naturally occurring long documents in real-world RAG systems, where irrelevant content can vary in topic, style, structure, and evidence density. Second, AnInter is trained mainly with moderate noise levels, $k = \{1, 2, 3, 4, 5\}$. The strong transfer to larger interference levels suggests that short-text anti-interference training is effective, but more noise schedules and document composition strategies remain to be explored. Third, our training experiments focus on the Qwen3-Embedding series. While the results across three model scales show consistent gains, further experiments are needed to verify whether the same training strategy generalizes to other retriever architectures. Finally, anti-interference training introduces a moderate drop on some standard retrieval benchmarks, indicating a trade-off between robustness to noisy positives and conventional semantic matching. Future work should study more realistic interference construction, more backbone families, and end-to-end RAG evaluation to examine whether improved anti-interference retrieval consistently benefits downstream generation.

References

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. [Llama-embed-nemotron-](#)

- 8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. **MS MARCO: A human generated machine reading comprehension dataset**. *arXiv preprint arXiv:1611.09268*.
- Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. **Overview of touché 2020: Argument retrieval**. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 384–395. Springer.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. **M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- 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*, pages 4171–4186. Association for Computational Linguistics.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. **CLIMATE-FEVER: A dataset for verification of real-world climate claims**. *arXiv preprint arXiv:2012.00614*.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, and 1 others. 2025. **MMTEB: Massive multilingual text embedding benchmark**. *arXiv preprint arXiv:2502.13595*.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. **ColPali: Efficient document retrieval with vision language models**. *arXiv preprint arXiv:2407.01449*.
- Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2024. **Late chunking: Contextual chunk embeddings using long-context embedding models**. *arXiv preprint arXiv:2409.04701*.
- Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2023. **Jina embeddings 2: 8192-token general-purpose text embeddings for long documents**. *arXiv preprint arXiv:2310.19923*.
- Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. **CQADupStack: A benchmark data set for community question-answering research**. In *Proceedings of the 20th Australasian Document Computing Symposium*, pages 3:1–3:8. Association for Computing Machinery.
- 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*.
- 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*.
- Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. **LLM maybe longLM: Self-extend LLM context window without tuning**. *arXiv preprint arXiv:2401.01325*.
- 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*, pages 6769–6781. Association for Computational Linguistics.
- 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. Association for Computing Machinery.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. **NV-Embed: Improved techniques for training LLMs as generalist embedding models**. *arXiv preprint arXiv:2405.17428*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. **Retrieval-augmented generation for knowledge-intensive NLP tasks**. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *arXiv preprint arXiv:2308.03281*.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [WWW’18 open challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.
- Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. [Generative representational instruction tuning](#). In *International Conference on Learning Representations*.
- Niklas Muennighoff, Nouamane Tazi, Loïc 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.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. [Large dual encoders are generalizable retrievers](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855. Association for Computational Linguistics.
- Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. [Nomic embed: Training a reproducible long context text embedder](#). *arXiv preprint arXiv:2402.01613*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.
- Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. [Benchmarking and building long-context retrieval models with LoCo and M2-BERT](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. [Efficient transformers: A survey](#). *ACM Computing Surveys*, 55(6):1–28.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: A large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819. 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*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. [TREC-COVID: Constructing a pandemic information retrieval test collection](#). *ACM SIGIR Forum*, 54(1):1–12.
- 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. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- 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.
- Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, and 9 others. 2020. [CORD-19: The COVID-19 open research dataset](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics.

- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.
- 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](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. 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. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.
- Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. [LongEmbed: Extending embedding models for long context retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

A AICB Results

Model	Ori.Hit@1	anIn@1	anIn@5	anIn@10	anIn@15	anIn@20	anIn@25	anIn@30	anIn@35	anIn@40	anIn@45	anIn@50	anIn@55	anIn@60	TanIn
Baseline															
llama-embed-nemotron-8B	83.6	50.3	8.2	4.1	2.6	1.8	1.4	1.2	0.9	0.8	0.6	0.5	0.4	0.4	5.6
NV-Embed-v2	95.4	68.8	14.1	4.7	2.2	1.7	1.4	0.9	0.5	0.3	0.3	0.3	0.4	0.9	7.4
E5-Mistral-7B-instruct	71.0	33.5	10.9	5.9	4.0	3.1	2.4	2.0	1.7	1.5	1.4	1.2	1.1	1.0	5.4
GTE-Qwen2-1.5B-instruct	53.7	23.3	7.7	4.2	2.8	2.2	1.8	1.5	1.3	1.1	1.0	0.9	0.8	0.8	3.8
GritLM-7B	52.6	21.9	5.8	3.0	2.1	1.6	1.3	1.1	1.0	0.8	0.8	0.7	0.6	0.6	3.2
instructor-xl	99.0	77.7	29.6	16.1	11.1	8.4	6.8	5.7	4.9	4.3	3.8	3.5	3.2	2.9	13.7
Qwen3-Embedding-0.6B	95.9	58.9	10.3	5.0	3.3	2.2	2.2	2.5	2.9	3.1	3.1	2.6	2.3	2.5	7.7
Qwen3-Embedding-4B	98.9	77.7	21.2	11.2	7.1	5.2	4.0	3.5	3.1	2.8	2.2	1.8	1.6	1.3	11.0
Qwen3-Embedding-8B	99.3	79.1	24.6	8.0	4.4	3.2	2.7	2.2	1.8	1.7	1.6	1.5	1.4	1.4	10.3
Our AnInter															
AnInter-0.6B	90.0	78.6	53.6	44.3	35.2	28.0	24.1	21.2	19.5	17.9	16.5	15.2	14.0	13.3	29.3
AnInter-4B	95.2	90.2	74.8	64.9	55.3	46.6	40.8	36.8	33.3	30.4	27.9	26.4	24.9	23.7	44.3
AnInter-8B	96.8	87.7	80.4	70.6	58.5	47.7	40.6	37.5	35.3	33.4	31.8	30.2	28.4	27.1	46.9
Δ -Best baseline	-2.5	+11.1	+50.8	+54.5	+47.4	+39.3	+34.0	+31.8	+30.4	+29.1	+28.0	+26.7	+25.2	+24.2	+33.2

Table 4: **Full results on AICB.** This table reports Ori.Hit@1, anIn@k at all thirteen test noise levels, and TanIn. The main paper reports representative anIn@k values for readability.

B Ablation Results

Train k	Ori.Hit@1	anIn@1	anIn@5	anIn@10	anIn@15	anIn@20	anIn@25	anIn@30	anIn@35	anIn@40	anIn@45	anIn@50	anIn@55	anIn@60	TanIn
Base	98.9	77.7	21.2	11.2	7.1	5.2	4.0	3.5	3.1	2.8	2.2	1.8	1.6	1.3	11.0
{1}	96.2	83.4	49.2	34.7	25.6	19.8	15.7	12.7	10.4	8.7	7.4	6.3	5.5	4.9	21.9
{1, 2, 3}	95.3	85.8	58.6	44.1	35.2	28.7	23.8	20.1	17.2	14.9	13.0	11.4	10.2	9.2	28.6
{1, 2, 3, 4, 5}	95.2	90.2	74.8	64.9	55.3	46.6	40.8	36.8	33.3	30.4	27.9	26.4	24.9	23.7	44.3
{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	89.3	85.9	61.8	48.7	40.5	35.1	31.6	28.2	25.4	23.2	21.3	19.7	18.4	17.2	35.2
{1, 2, 3, 4, 5}, no origin data	91.9	82.1	62.8	50.3	42.0	35.2	29.1	24.5	21.5	18.7	16.3	14.5	13.2	12.2	32.5
{1, 2, 3, 4, 5} + {6, 7, 8, 9, 10}	95.6	91.4	77.2	67.8	58.7	50.1	44.2	40.0	36.5	33.7	31.3	29.6	28.2	26.9	47.4

Table 5: **Full ablation results for training noise levels, origin anchoring, and progressive training.** All rows use Qwen3-Embedding-4B as the backbone. “Base” denotes the original Qwen3-Embedding-4B model without anti-interference training. The row “no origin data” removes origin query-document pairs from training. The row $k = \{1, 2, 3, 4, 5\} + \{6, 7, 8, 9, 10\}$ denotes progressive multi-stage training.

B-Ori.Hit@1	FT-Ori.Hit@1	anIn@1	anIn@5	anIn@10	anIn@15	anIn@20	anIn@25	anIn@30	anIn@35	anIn@40	anIn@45	anIn@50	anIn@55	anIn@60	TanIn
37.6	59.8	46.2	17.8	10.4	7.1	5.4	4.4	3.7	3.2	2.8	2.5	2.2	2.0	1.8	8.4
58.4	81.9	61.7	31.4	21.6	15.9	12.2	9.8	8.1	6.8	5.8	5.1	4.5	4.0	3.6	14.7
76.3	91.8	74.3	48.9	36.8	28.4	22.8	18.9	16.0	13.8	12.0	10.6	9.4	8.4	7.6	23.7
98.9	95.2	90.2	74.8	64.9	55.3	46.6	40.8	36.8	33.3	30.4	27.9	26.4	24.9	23.7	44.3

Table 6: **Effect of backbone origin retrieval ability.** All rows are fine-tuned from Qwen3-Embedding-4B with $k = \{1, 2, 3, 4, 5\}$ and the same LoRA setting as AnInter-4B. To test the importance of origin matching quality, we build training subsets with different backbone Ori.Hit@1 scores, denoted as “B-Ori.Hit@1”. “FT-Ori.Hit@1” reports the fine-tuned model’s accuracy on origin query-document pairs.

C Training Details

Hyperparameter	AnInter-0.6B	AnInter-4B	AnInter-8B
Backbone	Qwen3-Embedding-0.6B	Qwen3-Embedding-4B	Qwen3-Embedding-8B
Tuning method	LoRA	LoRA	LoRA
Training queries	3,656	3,656	3,656
Validation queries	522	522	522
Training noise levels	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}
Positive documents per query	21	21	21
Extra negatives per query	32	8	4
GPUs	8 RTX PRO 6000	8 RTX PRO 6000	8 RTX PRO 6000
Per-GPU query batch size	1	1	1
Gradient accumulation steps	1	1	1
Epochs	20	20	20
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Warmup ratio	0.1	0.1	0.1
Temperature τ	0.02	0.02	0.02
Weight decay	0.01	0.01	0.01
Max sequence length	10,000	10,000	10,000
Mixed precision	bf16	bf16	bf16
Gradient checkpointing	Yes	Yes	Yes
Max gradient norm	1.0	1.0	1.0
LoRA rank r	128	256	512
LoRA α	128	128	128
LoRA dropout	0.05	0.05	0.05
LoRA target modules	q, k, v, o projections and gate/up/down projections		
Random seed	42	42	42
GPU hours	15h	23h	31h

Table 7: **Training details of AnInter.** We report the main hyperparameters used to fine-tune the three AnInter models.