jina-reranker-v3: Last but Not Late Interaction for Listwise Document Reranking

Feng Wang[1] **Yuqing Li**[1,2] **Han Xiao**[1]

[1]Jina AI GmbH [2]University of Pittsburgh
Prinzessinnenstraße 19, 10969, Berlin, Germany
research@jina.ai

Abstract

jina-reranker-v3 is a 0.6B-parameter multilingual listwise reranker that introduces a novel *last but not late* interaction. Unlike late interaction models like ColBERT that encode documents separately before multi-vector matching, our approach applies causal attention between the query and all candidate documents in the same context window, enabling rich interactions before extracting contextual embeddings from each document's final token. The new model achieves state-of-the-art BEIR performance with 61.85 nDCG@10 while being significantly smaller than other models with comparable performance.

1 Introduction

Neural document retrieval faces a fundamental efficiency-effectiveness tradeoff. Cross-encoders achieve strong performance through joint query-document processing but require separate forward passes for each pair, while embedding models enable efficient similarity computation but lose fine-grained interaction signals. Recent models have attempted to bridge this gap through different interaction approaches. Late interaction models like ColBERT [Khattab and Zaharia, 2020] and their variants [Liu et al., 2024, Jha et al., 2024] separately encode queries and documents into multi-vector representations, then perform interaction through token-level similarity operations.

We introduce jina-reranker-v3, which features a novel *last but not late* interaction (LBNL) that takes a fundamentally different approach from existing methods. While late interaction models delay attention until after encoding documents separately, our method applies causal attention between the query and all documents within the context window, enabling cross-document interactions before extracting contextual embeddings from each document's ***last*** token. Unlike late interaction models that interact after encoding, we enable interactions during encoding—making our approach ***not late***. This "listwise" processing is not possible with separate encoding or bi-encoder approaches and represents our core innovation.

Evaluation shows jina-reranker-v3 achieves 61.85 nDCG@10 on BEIR [Thakur et al., 2021], representing the highest score among all evaluated rerankers and a 4.79% improvement over our previous jina-reranker-v2. The model excels particularly in multi-hop retrieval with HotpotQA reaching 78.58, fact verification achieving 94.01 on FEVER, competitive multilingual performance across 18 languages at 66.83 on MIRACL [Zhang et al., 2023a] and crosslingual retrieval with 67.92 Recall@10 on MKQA [Longpre et al., 2020] across 26 languages, and code retrieval reaching 70.64 on CoIR [Li et al., 2024].

Preprint.

2 Related Work

Document reranking approaches can be categorized by their interactions and learning objectives. Traditional learning-to-rank methods [Bruch et al., 2023] include pointwise approaches that predict relevance scores independently, pairwise methods like RankNet [Burges et al., 2005] that compare document pairs, and listwise techniques that optimize global ranking objectives. Cross-encoders like BERT-based rerankers [Nogueira and Cho, 2019] achieve strong performance through full query-document interaction but require separate forward passes for each pair, creating computational bottlenecks for large-scale retrieval. Recent comparative studies [Déjean et al., 2024] demonstrate that while LLM-based rerankers show impressive zero-shot capabilities, traditional cross-encoders remain highly competitive across diverse retrieval scenarios.

Late interaction models represent a significant approach that balances efficiency with expressiveness. ColBERT [Khattab and Zaharia, 2020] exemplifies this approach by independently encoding queries and documents into multi-vector representations, then computing similarity through MaxSim operations over token-level embeddings. This design enables pre-computation of document representations while preserving fine-grained matching signals. Recent developments have expanded this approach: analysis of matching mechanisms and token pruning strategies [Liu et al., 2024] provides theoretical foundations, LITE [Ji et al., 2024] introduces learnable late interactions, and Jina-ColBERT-v2 [Jha et al., 2024] extends the approach to multilingual settings. PyLate [Chaffin and Sourty, 2025] provides flexible frameworks for training and deployment of such models. The late chunking method [Günther et al., 2024] processes complete documents through transformers before applying chunking boundaries, extracting chunk-level embeddings that preserve contextual relationships. This approach demonstrates how leveraging broader document context can improve embedding quality, though it focuses primarily on retrieval rather than reranking applications.

LLM-powered reranker has emerged as a powerful family with diverse implementations. These approaches can be categorized into discriminative and generative methods. Generative approaches like RankGPT [Qin et al., 2023] prompt LLMs to generate ranked lists, leveraging their reasoning capabilities for relevance assessment, but typically require large models for competitive performance. Fine-tuning methods like RankVicuna [Pradeep et al., 2023] adapt existing models for relevance scoring tasks. Efficiency-focused innovations include FIRST [Reddy et al., 2024], which accelerates inference through single-token decoding, and PE-Rank [Qin et al., 2024], which leverages passage embeddings to reduce computational latency by $4.5\times$. Recent advances in training methodology include ERank [Cai et al., 2025], which combines supervised fine-tuning with reinforcement learning for improved ranking quality, and the Qwen3 Embedding series [Zhang et al., 2025], which demonstrates sophisticated multi-stage training pipelines. DeAR [Abdallah et al., 2025] introduces dual-stage reasoning with LLM distillation for enhanced cross-document analysis.

3 Model Architecture

`jina-reranker-v3` implements a new interaction that fundamentally differs from existing approaches. Built upon Qwen3-0.6B [Yang et al., 2025] with 28 transformer layers, 1024 hidden dimensions, 16 attention heads, and 131K token context capacity, our approach processes queries and multiple documents simultaneously within shared context windows. We add a lightweight MLP projector (1024→512→512 dimensions) to transform contextual representations into ranking-optimized embeddings. Table 5 provides complete architectural specifications.

3.1 Architecture

Figure 1 illustrates our architecture that addresses fundamental limitations in existing interactions. ColBERT [Khattab and Zaharia, 2020] achieves efficiency through separate encoding followed by multi-vector interaction, but cannot capture early query-document interactions during encoding or enable cross-document interactions within the attention mechanism.

Our LBNL approach enables causal self-attention interaction within the transformer architecture: instead of delaying interaction until after separate encoding as in late interaction models, we process all documents and the query simultaneously within shared context windows. This allows each document to attend to other documents and observe their content, enabling contextual embeddings that capture not just query-document relevance but also inter-document relationships and comparative

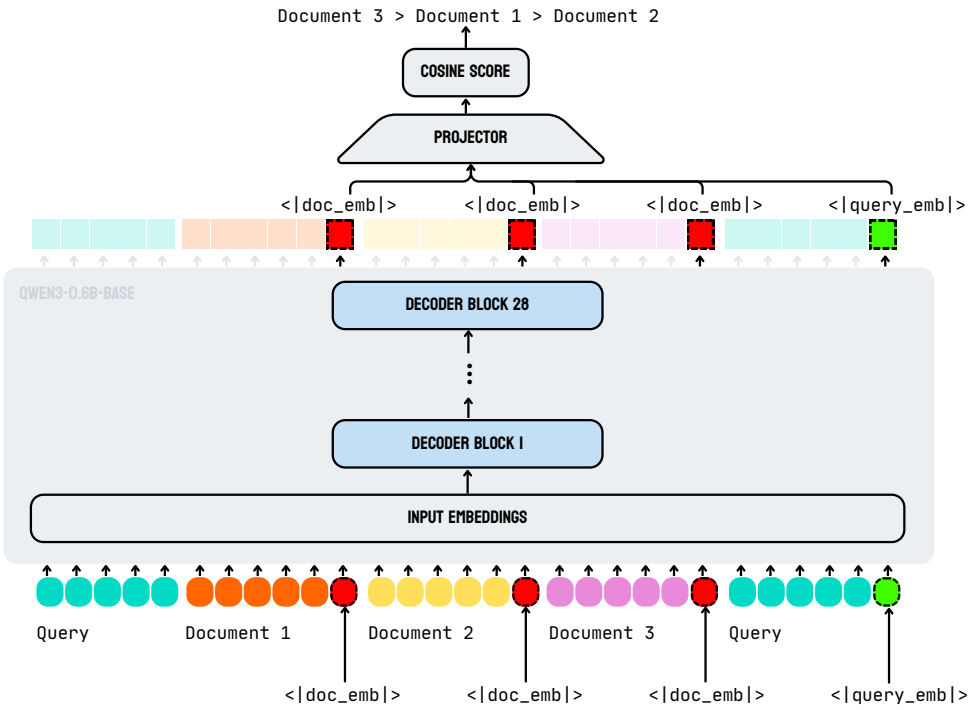

Figure 1: Architecture of `jina-reranker-v3` showing the transformer backbone with special token positions for embedding extraction. The model processes multiple documents and query in one context window, extracting contextual embeddings at designated token positions for similarity computation.

context. Such cross-document interactions are impossible in separate encoding approaches and represent a fundamental advancement in `jina-reranker-v3` architecture.

We extract contextual embeddings at designated special token positions: $\tilde{\mathbf{q}} = \mathbf{H}_{t_q}$ and $\tilde{\mathbf{d}}_i = \mathbf{H}_{t_i}$ where t_q and t_i are positions of the special tokens and $\mathbf{H}$ represents the transformer's final layer hidden states after causal self-attention. These embeddings capture both local document semantics and global cross-document context through the shared attention mechanism, enabling rich inter-document interactions unavailable in separate encoding approaches.

A two-layer projection network with ReLU activation maps the 1024-dimensional hidden states to 256-dimensional embedding space: $\mathbf{q} = P_\phi(\tilde{\mathbf{q}})$ and $\mathbf{d}_i = P_\phi(\tilde{\mathbf{d}}_i)$. Relevance scores are computed via cosine similarity: $s_i = \cos(\mathbf{q}, \mathbf{d}_i)$. This architecture combines the expressiveness of joint encoding with efficient similarity computation.

For document collections exceeding the 131K token context limit, we process documents in batches of up to 64 documents per forward pass, with query embeddings maintained consistently across batches to ensure ranking coherence.

3.2 Prompt Template

`jina-reranker-v3` processes structured prompts following Qwen3's instruction format with `system`/`user`/`assistant` roles to leverage existing instruction-following capabilities. As shown in Table 1, the system prompt establishes a search relevance expert persona, while the user prompt provides clear ranking instructions with dual query placement.

The template strategically places the query both at the beginning for instructions and at the end for final attention, sandwiching all documents in between. This design enables the final query position to attend to all preceding documents through causal attention while maintaining clear task instructions. Special tokens `<|doc_emb|>` after each document and `<|query_emb|>` after the final query mark specific positions for embedding extraction from transformer hidden states.

Prompt Template

```
<|im_start|>system
You are a search relevance expert who can determine
a ranking of passages based on their relevance to the query.
<|im_end|>
<|im_start|>user
I will provide you with k passages, each indicated by a numerical identifier.
Rank the passages based on their relevance to query: [QUERY]

<passage id="1">
[DOCUMENT_1]<|doc_emb|>
</passage>
<passage id="2">
[DOCUMENT_2]<|doc_emb|>
</passage>
...
<passage id="k">
[DOCUMENT_k]<|doc_emb|>
</passage>

<query>
[QUERY]<|query_emb|>
</query>
<|im_end|>
```

Table 1: Complete prompt template structure used by `jina-reranker-v3`. Special tokens `<|doc_emb|>` and `<|query_emb|>` mark positions for embedding extraction from transformer hidden states.

4 Training

4.1 Loss Functions

`jina-reranker-v3` employs a comprehensive multi-objective training approach combining InfoNCE loss with specialized auxiliary losses to optimize ranking performance across diverse domains.

The core training objective integrates multiple loss components, each addressing distinct aspects of the ranking problem:

$$\ell = \ell_{\text{rank}} + 0.45 \cdot \ell_{\text{disperse}} + 0.85 \cdot \ell_{\text{dual}} + 0.85 \cdot \ell_{\text{similar}} \tag{1}$$

The primary component is the InfoNCE loss ℓ_{rank} [van den Oord et al., 2019], which generates the core ranking signal through contrastive learning with hard negatives:

$$\ell_{\text{rank}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\mathbf{q}_i, \mathbf{d}_i^+)/\tau}}{Z_i} \quad \text{where} \quad Z_i = e^{s(\mathbf{q}_i, \mathbf{d}_i^+)/\tau} + \sum_{k=1}^{K} e^{s(\mathbf{q}_i, \mathbf{d}_{i,k}^-)/\tau} \tag{2}$$

Here, $\mathbf{q}_i$ denotes the query embedding, $\mathbf{d}_i^+$ represents the positive document embedding, $\mathbf{d}_{i,k}^-$ denotes one of K negative document embeddings, $s(\cdot,\cdot)$ is the cosine similarity function, τ is the temperature parameter, and N is the batch size.

To prevent representation collapse, we incorporate the dispersive loss ℓ_{disperse} [Wang et al., 2024], which enhances embedding diversity by maximizing the average pairwise cosine distance between document embeddings:

$$\ell_{\text{disperse}} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{K}\sum_{k=1}^{K} \left(e^{s(\mathbf{d}_i^+, \mathbf{d}_{i,k}^-)/\tau} + \sum_{k'=k}^{K-1} e^{s(\mathbf{d}_{i,k}^-, \mathbf{d}_{i,k'+1}^-)/\tau} \right) \tag{3}$$

The dual matching loss ℓ_{dual}[1] follows the same formulation as Eq. 2 but computes the query embedding from the query tokens at the sequence start. This enforces bidirectional consistency between query-to-document and document-to-query similarity scores, enhancing ranking robustness.

Finally, the similarity loss ℓ_{similar} [Huang et al., 2024] maintains semantic coherence at the document level. For each document in the input set, we create an augmented duplicate $\mathbf{d}_i^*$ through text augmentation techniques. The loss then treats the original document and its augmented version as a positive pair, while other documents serve as negatives. This encourages consistent embedding representations for semantically equivalent documents, even when their surface forms differ due to augmentation.

4.2 Multi-Stage Training

The training methodology follows a progressive three-stage approach designed for systematic complexity scaling:

Stage 1: Foundation Specialization. Starting from pretrained `Qwen3-0.6B`, we simultaneously train domain-specific configurations using LoRA fine-tuning with r=16 and α=32 targeting all attention and FFN layers while freezing the backbone. The model processes training sequences containing 16 documents per query (one positive and 15 negative examples), with each document truncated or padded to 768 tokens, yielding a maximal total sequence length of 12,288 tokens. Training data is drawn from diverse datasets including BGE-M3 [Chen et al., 2024] for multilingual coverage across 15 languages, Cornstack [Suresh et al., 2025] for code retrieval, as well as specialized datasets for biomedical [Xu et al., 2024] and instruction following [Weller et al., 2024] configurations.

Stage 2: Context and Hard Negative Mining. This stage combines context extension and comprehensive robustness optimization. Context extension is implemented in two ways: (1) extending individual document length to 8,192 tokens through datasets like MLDR [Chen et al., 2024] for long-document understanding, and (2) increasing the number of negative documents from 15 to 45 per query while maintaining the total sequence length under 131K tokens. Simultaneously, cross-system hard negative mining ensures robustness through specialized optimizations including *jina-en-v2* for English performance, *miracl-v2* for multilingual retrieval, *cornstack-v2* for code understanding, and *context-chunk-v3* for long-document processing. Training systematically mines hard negatives across multiple retrieval systems including BGE, Jina, GTE, and E5-Large with up to 25 negatives per query and very low temperature of 0.05, using key datasets including MS-MARCO Campos et al. [2016], mMARCO Bonifacio et al. [2021], and domain-specific synthetic question-answer pairs.

Stage 3: Model Ensemble and Optimization. The final stage combines multiple specialized models trained in previous stages through linear model merging. Each domain-specific model contributes weighted expertise, with merge weights ranging from 0.25 to 0.65 based on domain importance and performance. This approach enables the final model to leverage diverse domain knowledge while maintaining architectural efficiency.

Detailed hyperparameter evolution across stages demonstrates multi-objective optimization with stage-tailored configurations (see Appendix 6). Foundation stages use aggressive learning rates of 5e-5 with substantial negative sampling of 15 negatives. Context scaling stages reduce batch sizes dramatically from 60 to 6 to accommodate 8K sequences while employing conservative learning rates of 6e-6. Loss weight adaptation varies across different domain specializations, with dispersive loss typically set to 0.45, dual-matching loss ranging from 0.65 to 0.85, and similarity loss stabilizing around 0.75 to 0.85 depending on the specific domain requirements.

5 Evaluation

5.1 Experimental Setup

Our evaluation spans four challenging benchmarks that test different aspects of ranking capability. BEIR Thakur et al. [2021] represents the gold standard for English retrieval evaluation, encompassing 13 heterogeneous tasks from question answering on Natural Questions to fact verification on

[1]During training, the special token `<|query_emb|>` is inserted at the end of the query at the beginning of the input sequence.

FEVER, testing the model's ability to generalize across domains without task-specific optimization. MIRACL Zhang et al. [2023b] pushes multilingual boundaries with 18 languages spanning diverse linguistic families, from Arabic and Chinese to Finnish and Thai, requiring deep cross-lingual understanding. MKQA Longpre et al. [2021] specifically challenges cross-lingual question answering capabilities, while CoIR Li et al. [2025] focuses on the specialized domain of code retrieval, where semantic understanding of programming constructs becomes crucial.

The first-stage dense retriever is `jina-embeddings-v3`, providing the foundation top-100 candidates that all rerankers process. Second-stage rerankers encompass our previous `jina-reranker-v2`, the multilingual `bge-reranker-v2-m3`, the mxbai-rerank variants at different scales, and `Qwen3-Reranker-0.6B` and `Qwen3-Reranker-4B` models.

5.2 Overall Performance Across Benchmarks

Table 2 demonstrates `jina-reranker-v3`'s exceptional performance density across diverse evaluation scenarios. On BEIR, our model achieves the highest score among all rerankers at 61.85, establishing new state-of-the-art performance for English retrieval. This represents a 4.79% improvement over our previous `jina-reranker-v2` at 57.06, directly attributable to LBNL interaction mechanism where query and document embeddings are extracted from shared forward passes rather than separate encoding pipelines.

Parameter efficiency analysis reveals striking advantages compared to larger alternatives. Against the 1.5B parameter `mxbai-rerank-large-v2`, `jina-reranker-v3` achieves superior BEIR performance with 61.85 versus 61.44 using $2.5\times$ fewer parameters, while providing specialized domain coverage unavailable in competing models reaching 70.64 on CoIR. This efficiency derives from architectural innovations: Qwen3's optimized transformer backbone combined with our specialized 512-dimensional projector network that concentrates ranking signals without requiring massive parameter scaling.

Multilingual evaluation reveals strong cross-lingual capabilities despite the model's compact architecture. The 66.83 score on MIRACL, while 2.49 points below the multilingual-specialized `bge-reranker-v2-m3` at 69.32, demonstrates effective knowledge transfer from our progressive training methodology. The 67.92 MKQA performance closely approaches `jina-reranker-m0`'s 68.19, indicating that architectural sophistication can partially offset parameter differences in multilingual scenarios.

Models	**# Param**	**BEIR**	**MIRACL**	**MKQA**	**CoIR**
First-stage Retriever					
`jina-embeddings-v3`	0.5B	55.81	58.90	65.63	-
`jina-code-embeddings-0.5b`	-	-	-	-	**73.94**
Second-stage Reranker					
`jina-reranker-v3`	0.6B	**61.85**	66.83	67.92	70.64
`jina-reranker-v2`	0.3B	57.06	63.65	67.90	58.35
`jina-reranker-m0`	2.4B	58.95	66.75	68.19	66.89
`bge-reranker-v2-m3`	0.6B	56.51	**69.32**	67.88	36.28
`mxbai-rerank-base-v2`	0.5B	58.40	55.32	64.24	65.71
`mxbai-rerank-large-v2`	1.5B	61.44	57.94	67.06	70.87
`Qwen3-Reranker-0.6B`	0.6B	56.28	57.70	65.34	65.18
`Qwen3-Reranker-4B`	4.0B	61.16	67.52	**69.25**	73.91

Table 2: Evaluation results for all rerankers. All scores are from our runs based on the top-100 retrieval results from the first row. For MKQA, we used Recall@10; for all other benchmarks, we used NDCG@10.

5.3 English Retrieval Performance on BEIR

Table 3 provides granular analysis across BEIR's heterogeneous tasks, revealing specific architectural advantages. The model achieves consistent excellence across diverse reasoning tasks, with

particularly strong performance on complex multi-hop reasoning reaching 78.58 on HotpotQA and fact verification achieving 94.01 on FEVER. These results highlight how LBNL interaction enables sophisticated query-document self-attention during encoding, capturing evidence relationships that separate encoding approaches miss.

Within the same scale category, `jina-reranker-v3` reveals significant advantages. Against `bge-reranker-v2-m3` with the same 0.6B parameters, `jina-reranker-v3` delivers a substantial 5.34% improvement from 56.51 to 61.85, demonstrating architectural innovation over simple parameter scaling. The specialized 512-dimensional projector network effectively concentrates ranking signals while preserving contextual representations from the Qwen3 backbone. Remarkably, our model surpasses `mxbai-rerank-large-v2`'s 61.44 performance while using $2.5\times$ fewer parameters, establishing that sophisticated architecture can surpass brute-force scaling approaches.

Since `jina-reranker-v3` processes all documents simultaneously in a listwise manner within shared context windows, we investigate the sensitivity to document ordering. We evaluate three variants: documents ordered by descending relevance scores (D), ascending scores (A), and random permutation (R). The results show modest variations across orderings, with random ordering (R) achieving the highest average of 62.24, followed by descending (D) at 61.85 and ascending (A) at 61.45. While the differences are not conclusive, this analysis reveals that the LBNL interaction maintains relatively stable performance across different input orderings, suggesting robust self-attention mechanisms that can effectively process documents regardless of their initial arrangement.

The model's dominance extends particularly to question-answering scenarios, where Natural Questions achieves 74.28 and argumentative retrieval on ArguAna reaches 73.43, showcasing the benefit of contextual embeddings. These tasks require understanding complex query intent and matching it against nuanced document semantics, precisely the scenario where our dual embedding extraction approach provides maximum advantage over traditional cross-encoder scoring.

Models	**Size**	**Avg.**	**TC**	**NFC**	**NQ**	**HQA**	**FQA**	**AA**	**TCH**	**DBP**	**SD**	**FVR**	**CFV**	**SF**	**QRA**
First-stage Retriever															
`jina-embeddings-v3`	0.5B	55.81	77.81	36.65	64.31	64.63	47.47	54.31	26.55	41.07	19.91	89.00	42.33	72.4	89.06
Second-stage Reranker															
`jina-reranker-v3` (D)	0.6B	61.85	84.75	37.66	**74.28**	**78.58**	49.16	73.43	32.24	47.98	23.23	**94.01**	41.63	76.51	**90.63**
`jina-reranker-v3` (A)	0.6B	61.45	85.90	39.14	72.34	77.48	50.99	69.36	29.73	48.30	23.90	93.46	41.72	76.75	89.73
`jina-reranker-v3` (R)	0.6B	**62.24**	86.59	38.92	72.90	78.03	51.81	74.12	30.12	48.37	24.26	93.84	43.05	76.84	90.24
`jina-reranker-m0`	2.4B	58.95	84.17	41.03	72.25	76.99	51.62	40.69	31.79	49.34	22.91	91.14	36.42	**79.94**	88.01
`jina-colbert-v2`	0.6B	54.49	81.94	35.88	66.01	74.36	43.62	35.46	29.11	47.14	19.40	87.92	29.20	70.13	88.25
`jina-reranker-v2`	0.3B	57.06	80.53	37.17	67.39	76.17	46.48	39.28	32.35	47.81	20.03	93.02	37.17	76.50	87.83
`bge-reranker-v2-m3`	0.6B	56.51	82.19	34.33	69.52	77.89	45.45	36.21	33.12	46.72	17.79	91.03	38.69	72.64	89.10
`mxbai-rerank-base-v2`	0.5B	58.40	82.75	37.57	67.74	77.35	47.33	47.33	30.71	48.00	18.09	93.30	42.93	77.76	88.33
`mxbai-rerank-large-v2`	1.5B	61.44	81.51	37.76	72.46	78.10	**52.75**	**74.55**	29.81	49.07	18.58	93.94	42.03	78.86	89.36
`Qwen3-Reranker-0.6B`	0.6B	56.28	87.08	38.37	56.54	74.41	43.45	56.53	27.26	43.54	20.98	86.19	44.11	74.89	78.32
`Qwen3-Reranker-4B`	4.0B	61.16	**87.08**	**41.56**	69.06	77.03	52.29	58.82	**33.73**	**50.81**	**26.01**	87.80	**47.59**	78.41	84.83

Table 3: Performances of different rerankers (nDCG@10 in %) on BEIR. Top-100 retrieval results from `jina-embeddings-v3` are passed as input. The best results are marked in bold. Avg. represents the averaged result of the 13 BEIR datasets. For `jina-reranker-v3`, (D)/(A)/(R) denote document ordering variants: Descending, Ascending, and Random relevance score ordering, respectively.

5.4 Multilingual Performance on MIRACL

MIRACL evaluation across 18 diverse languages demonstrates `jina-reranker-v3`'s cross-lingual consistency despite its compact architecture. The 66.50 average performance reveals sophisticated multilingual understanding, with particularly strong results in morphologically complex languages like Arabic achieving 78.69 and challenging contexts like Thai reaching 81.06. These results reflect the effectiveness of our progressive multilingual training strategy, where architectural advantages help compensate for reduced multilingual specialization.

Perhaps most significantly, `jina-reranker-v3` exhibits minimal performance degradation across linguistic families, from Indo-European languages like Russian at 65.20 to Sino-Tibetan languages like Thai at 81.06. This consistency stems from our progressive multilingual training strategy that incorporates diverse datasets including MIRACL, mMARCO, and domain-specific multilingual

corpora during the three-stage training progression. The architectural advantage becomes particularly evident in Korean achieving 73.83, where the model's LBNL interaction enables effective handling of complex agglutinative morphology that traditional cross-encoders struggle to process efficiently.

Compared to `bge-reranker-v2-m3`'s dedicated multilingual optimization averaging 69.32, `jina-reranker-v3` accepts a 2.82-point gap while achieving superior English performance and maintaining architectural efficiency. This trade-off reflects our design philosophy: contextual embedding extraction provides competitive multilingual capabilities without massive multilingual scaling, creating an optimal balance for applications requiring both English excellence and cross-lingual competency.

Models	Avg.	AR	BN	EN	ES	FA	FI	FR	HI	ID	JA	KO	RU	SW	TE	TH	ZH	DE	YO
First-stage Retriever																			
`jina-embeddings-v3`	58.90	71.53	69.86	48.37	46.91	54.13	71.15	50.90	55.05	47.83	56.46	64.76	55.63	54.07	70.48	73.56	55.29	49.18	65.01
Second-stage Reranker																			
`jina-reranker-v3`	66.83	78.85	79.47	**59.45**	54.57	57.70	76.03	55.74	61.52	57.43	65.94	73.60	65.50	64.54	74.53	81.57	**65.60**	56.74	74.07
`jina-reranker-m0`	66.75	79.78	78.01	59.21	53.55	58.90	70.00	56.66	62.83	54.92	66.51	72.86	67.37	59.04	70.19	80.37	64.51	58.50	80.44
`jina-reranker-v2`	63.65	72.50	79.42	46.66	51.54	57.81	73.05	50.90	60.94	56.66	59.15	72.60	53.43	66.47	74.62	77.75	62.49	53.06	76.69
`bge-reranker-v2-m3`	**69.32**	**80.51**	**81.85**	57.67	**57.64**	**61.92**	**80.38**	**59.60**	67.66	**58.86**	**67.37**	**75.14**	**67.61**	**68.92**	**76.69**	**82.29**	64.46	**58.32**	80.85
`mxbai-rerank-base-v2`	55.32	71.08	58.21	56.61	48.89	46.59	64.92	50.47	44.75	49.48	57.99	64.88	54.16	48.40	55.15	72.71	58.44	20.33	72.66
`mxbai-rerank-large-v2`	57.94	71.38	63.48	57.55	49.14	48.38	66.70	51.61	45.12	49.05	56.61	64.98	54.80	51.79	62.41	74.51	62.29	38.66	74.42
`Qwen3-Reranker-0.6B`	56.16	67.44	66.67	50.91	45.77	52.07	65.50	43.28	60.36	49.66	51.56	61.03	48.88	46.72	69.86	72.95	45.14	43.00	70.04
`Qwen3-Reranker-4B`	67.52	78.32	81.51	59.37	53.07	61.63	78.70	55.02	**68.71**	54.90	65.32	71.80	64.66	66.50	75.60	82.00	59.35	57.56	**81.39**

Table 4: Multilingual retrieval performance on the MIRACL (measured by nDCG@10).

6 Conclusion

We present `jina-reranker-v3`, a 0.6B-parameter multilingual listwise reranker that introduces *last but not late* interaction for efficient document reranking. Our approach enables cross-document interactions during encoding by processing queries and multiple documents simultaneously within shared context windows, then extracting contextual embeddings from designated special token positions. By adapting long-context generative LLMs into a discriminative model, `jina-reranker-v3` bridges the efficiency-effectiveness gap while maintaining significant computational advantages over generative rerankers. Future work includes studying the robustness of ranking against prompt injections and deduplication within the context window using submodularity optimization.

References

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. DeAR: Dual-Stage Document Reranking with Reasoning Agents via LLM Distillation. *arXiv preprint arXiv:2508.16998*, 2025. URL `https://arxiv.org/abs/2508.16998`. Accepted at EMNLP Findings 2025.

L. Bonifacio, Israel Campiotti, R. Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021. URL `https://arxiv.org/abs/2108.13897`.

Sebastian Bruch, C. Lucchese, and F. M. Nardini. Efficient and effective tree-based and neural learning to rank. *Foundations and Trends in Information Retrieval*, 2023. doi: 10.1561/1500000071.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, 2005. doi: 10.1145/1102351.1102363. URL `https://doi.org/10.1145/1102351.1102363`.

Yuzheng Cai, Yanzhao Zhang, Dingkun Long, Mingxin Li, Pengjun Xie, and Weiguo Zheng. ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking. *arXiv preprint arXiv:2509.00520*, 2025. URL `https://arxiv.org/abs/2509.00520`.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, 2016.

Antoine Chaffin and Raphael Sourty. Pylate: Flexible training and retrieval for late interaction models. *arXiv.org*, 2025. doi: 10.48550/arXiv.2508.03555.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. A Thorough Comparison of Cross-Encoders and LLMs for Reranking SPLADE. *arXiv preprint arXiv:2403.10407*, 2024. URL `https://doi.org/10.48550/arXiv.2403.10407`.

Michael Günther, Isabelle Mohr, Bo Wang, and Han Xiao. Late Chunking: Contextual Chunk Embeddings Using Long-Context Embedding Models. *arXiv preprint arXiv:2409.04701*, 2024. URL `https://arxiv.org/abs/2409.04701`. Submitted to ICLR 2025.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2024. doi: 10.1109/TASLP.2024.3402087.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *arXiv preprint arXiv:2408.16672*, 2024.

Ziwei Ji, Himanshu Jain, Andreas Veit, Sashank J. Reddi, Sadeep Jayasumana, Ankit Singh Rawat, Aditya Krishna Menon, Felix X. Yu, and Sanjiv Kumar. Efficient Document Ranking with Learnable Late Interactions. *arXiv preprint arXiv:2406.17968*, 2024. URL `https://arxiv.org/abs/2406.17968`.

Omar Khattab and Matei Zaharia. 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, 2020. doi: 10.1145/3397271.3401075. URL `https://arxiv.org/abs/2004.12832`.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Yichun Yin, Hao Zhang, Yong Liu, Yasheng Wang, and Ruiming Tang. Coir: A comprehensive benchmark for code information retrieval models, 2024. URL `https://arxiv.org/abs/2407.02883`.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. Coir: A comprehensive benchmark for code information retrieval models. In *ACL*, pages 22074–22091, 2025. URL `https://aclanthology.org/2025.acl-long.1072/`.

Qi Liu, Gang Guo, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, and Zhao Cao. An analysis on matching mechanisms and token pruning for late-interaction models. *ACM Trans. Inf. Syst.*, 2024. doi: 10.1145/3639818.

Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering, 2020. URL `https://arxiv.org/pdf/2007.15207.pdf`.

Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021. doi: 10.1162/TACL_A_00433. URL `https://arxiv.org/abs/2007.15207`.

Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. URL `https://arxiv.org/abs/1901.04085`.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *arXiv preprint arXiv:2309.15088*, 2023. URL https://arxiv.org/abs/2309.15088.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 13053–13074, 2023. URL https://arxiv.org/abs/2306.17563.

Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xinyu Zhang, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. Leveraging Passage Embeddings for Efficient Listwise Reranking with Large Language Models. *arXiv preprint arXiv:2406.14848*, 2024. URL https://arxiv.org/abs/2406.14848.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md. Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. FIRST: Faster Improved Listwise Reranking with Single Token Decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, 2024. doi: 10.18653/v1/2024.emnlp-main.491. URL https://arxiv.org/abs/2406.15657.

Tarun Suresh, Revanth Gangi Reddy, Yifei Xu, Zach Nussbaum, Andriy Mulyar, Brandon Duderstadt, and Heng Ji. Cornstack: High-quality contrastive data for better code retrieval and reranking, 2025.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2019. URL https://arxiv.org/abs/1807.03748.

Lu Wang, Chao Du, Pu Zhao, Chuan Luo, Zhangchi Zhu, Bo Qiao, Wei Zhang, Qingwei Lin, S. Rajmohan, Dongmei Zhang, and Qi Zhang. Contrastive learning with negative sampling correction. *arXiv preprint arXiv:2401.08690*, 2024. doi: 10.48550/arXiv.2401.08690. URL https://arxiv.org/abs/2401.08690.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions, 2024.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D. Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. Bmretriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 09 2023a. ISSN 2307-387X. doi: 10.1162/tacl_a_00595. URL https://doi.org/10.1162/tacl_a_00595.

Xinyu Crystina Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy J. Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 2023b. doi: 10.1162/tacl_a_00595.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Guangwei Xu, and Pengjun Xie. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A Model Configuration and Training Details

Parameter	Value
Total Parameters	0.6B
Non-Embedding Parameters	0.44B
Hidden Size	1,024
Number of Layers	28
Attention Heads (Q/KV)	16/8 (GQA)
Context Length	131,072
Effective Sequence Length	8,192
Projector Architecture	1024→512→512
Projector Activation	ReLU

Table 5: Model architecture configuration for `jina-reranker-v3`.

Hyperparameter	Stage 1 Foundation	Stage 2 Context & Hard Mining	Stage 3 Model Ensemble
Learning Rate	5e-5	[5e-5, 6e-6]	-
Batch Size (per device)	60	[6, 60]	-
Max Sequence Length	[768, 2048]	[2048, 8192]	-
Max Query Length	-	[256, 512]	-
Max Doc Length	-	[512, 2048]	-
Number of Negatives	15	[9, 25]	-
In-batch Negatives	3	[0, 3]	-
Temperature	0.25	[0.05, 0.25]	-
Training Mode	LoRA	LoRA/Full	Linear Merging
LoRA Rank	16	16	-
Word Embeddings	Tuned	Frozen/Tuned	-
Backbone	Frozen	Frozen/Tuned	-
Dispersive Loss α	0.45	[0.25, 0.45]	-
Dual Matching α	0.85	[0.65, 0.85]	-
Similarity Loss α	0.85	[0.75, 0.85]	-

Table 6: Multi-stage supervised fine-tuning hyperparameters showing ranges across 47 training configurations.