Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models

Yanzhao Zhang* Mingxin Li* Dingkun Long* Xin Zhang*
Huan Lin Baosong Yang Pengjun Xie An Yang
Dayiheng Liu Junyang Lin Fei Huang Jingren Zhou
Tongyi Lab Alibaba Group

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3-Embedding

Abstract

In this work, we introduce the Qwen3 Embedding series, a significant advancement over its predecessor, the GTE-Qwen series, in text embedding and reranking capabilities, built upon the Qwen3 foundation models. Leveraging the Qwen3 LLMs' robust capabilities in multilingual text understanding and generation, our innovative multi-stage training pipeline combines large-scale unsupervised pre-training with supervised fine-tuning on high-quality datasets. Effective model merging strategies further ensure the robustness and adaptability of the Qwen3 Embedding series. During the training process, the Qwen3 LLMs serve not only as backbone models but also play a crucial role in synthesizing high-quality, rich, and diverse training data across multiple domains and languages, thus enhancing the training pipeline. The Qwen3 Embedding series offers a spectrum of model sizes (0.6B, 4B, 8B) for both embedding and reranking tasks, addressing diverse deployment scenarios where users can optimize for either efficiency or effectiveness. Empirical evaluations demonstrate that the Qwen3 Embedding series achieves state-of-the-art results across diverse benchmarks. Notably, it excels on the multilingual evaluation benchmark MTEB for text embedding, as well as in various retrieval tasks, including code retrieval, cross-lingual retrieval and multilingual retrieval. To facilitate reproducibility and promote community-driven research and development, the Qwen3 Embedding models are publicly available under the Apache 2.0 license.

1 Introduction

Text embedding and reranking are fundamental components in numerous natural language processing and information retrieval applications, including web search, question answering, recommendation systems, and beyond (Karpukhin et al., 2020; Huang et al., 2020; Zhao et al., 2023; 2024). High-quality embeddings enable models to capture semantic relationships between texts, while effective reranking mechanisms ensure that the most relevant results are prioritized. Recently, emerging application paradigms such as Retrieval-Augmented Generation (RAG) and agent systems, driven by the advancement of large language models (e.g., Qwen3 (Yang et al., 2025), GPT-4o (Hurst et al., 2024)), have introduced new requirements and challenges for text embedding and reranking, both in terms of model training paradigms and application scenarios. Despite significant advancements, training embedding and reranking models that perform well in scalability, contextual understanding, and alignment with specific downstream tasks remains challenging.

The emergence of large language models (LLMs) has significantly advanced the development of text embedding and reranking models. Prior to the introduction of LLMs, the predominant approach

* Equal contribution

QWEN3嵌入：通过基础模型推进文本嵌入和重新固定

Yanzhao Zhang* mingxin li* dingkun long* Xin Zhang* Huan Lin B
aosong Yang Pengjun Xie an Yang Dayiheng Liu Junyang Junyang li
n fei lin fei lin fei lin fei huang jingren zhou jingren zhou tongyi zhou
tongyi实验室

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3-Embedding

抽象的

在这项工作中，我们介绍了QWEN3嵌入式系列，这是对其前身GTE-QWEN系列的重大进步，该系列构建了Qwen3 Foundation Models构建的文本嵌入和重新研究。利用QWEN3 LLMS在多语言文本理解和产生中的强大功能，我们创新的多阶段培训管道将大规模无监督的预训练与高质量数据集中的监督微调相结合。有效的模型合并策略进一步确保了QWEN3嵌入序列的鲁棒性和适应性。在培训过程中，QWEN3 LLM不仅用作骨干模型，而且在跨多个领域和语言的高质量，丰富和多样化的培训数据中发挥着至关重要的作用，从而增强了培训管道。 QWEN3嵌入式系列提供了嵌入和重新管理任务的一系列模型尺寸（0.6b，4b，8b），以解决用户可以优化效率或有效性的各种部署方案。经验评估表明，QWEN3嵌入序列可在不同的基准中实现最先进的结果。值得注意的是，它在多种语言评估基准MTEB上符合文本嵌入以及各种检索任务，包括代码检索，跨语性检索和多语言检索。为了促进可重复性并促进社区驱动的研发，QWEN3嵌入模型可在Apache 2.0许可下公开使用。

1简介

文本嵌入和重读是许多自然语言进程和信息检索应用中的基本组成部分，包括Web搜索，问答，推荐系统以及其他（Karpukhin等，2020; Huang等，2020; Zhao et al。，2023; 2023; 2024; 2024; 2024; 2024; 2024; 2024）。高质量的嵌入使模型能够捕获文本之间的语义关系，而有效的重新疗法机制可确保优先考虑最相关的结果。 Recently, emerging application paradigms such as Retrieval-Augmented Generation (RAG) and agent sys- tems, driven by the advancement of large language models (e.g., Qwen3 (Yang et al., 2025), GPT-4o (Hurst et al., 2024)), have introduced new requirements and challenges for text embedding and reranking, both in terms of model training paradigms and application scenarios.尽管取得了重大进步，但在可扩展性，上下文理解以及与特定下游任务保持一致性方面表现良好的培训嵌入和重新融合模型仍然具有挑战性。

大型语言模型（LLM）的出现已经显着推动了文本嵌入和重新依给模型的发展。在引入LLM之前，主要方法

* Equal contribution

involved using encoder-only pretrained language models like BERT as the foundational model for training (Reimers & Gurevych, 2019). The richer world knowledge, text understanding, and reasoning abilities inherent in LLMs have led to further enhancements in models trained on these architectures. Additionally, there has been considerable research facilitating the integration of LLMs into processes such as training data synthesis and quality data filtering (Wang et al., 2024; Lee et al., 2024; 2025b). The fundamental characteristics of LLMs have also inspired the introduction of new training paradigms. For instance, during the embedding model training process, incorporating differentiated tasks across aspects such as instruction type, domain, and language has yielded improved performance in downstream tasks (Su et al., 2023). Similarly, for reranking model training, advancements have been realized through both zero-shot methods based on user prompts and approaches combining supervised fine-tuning (Ma et al., 2023; Pradeep et al., 2023; Zhang et al., 2024a; Zhuang et al., 2024).

In this work, we introduce the Qwen3 Embedding series models, which are constructed on top of the Qwen3 foundation models. The Qwen3 foundation has simultaneously released base and instruct model versions, and we exploit the robust multilingual text understanding and generation capabilities of these models to fully realize their potential in training embedding and reranking models. To train the embedding models, we implement a multi-stage training pipeline that involves large-scale unsupervised pre-training followed by supervised fine tuning on high-quality datasets. We also employ model merging with various model checkpoints to enhance robustness and generalization. The Qwen3 instruct model allows for efficient synthesis of a vast, high-quality, multilingual, and multi-task text relevance dataset. This synthetic data is utilized in the initial unsupervised training stage, while a subset of high-quality, small-scale data is selected for the second stage of supervised training. For the reranking models, we adopt a two-stage training scheme in a similar manner, consisting of high-quality supervised fine tuning and a model merging stage. Based on different sizes of the Qwen3 backbone models (including 0.6B, 4B, and 8B), we ultimately trained three text embedding models and three text reranking models. To facilitate their application in downstream tasks, the Qwen3 Embedding series supports several practical features, such as flexible dimension representation for embedding models and customizable instructions for both embedding and reranking models.

We evaluate the Qwen3 Embedding series across a comprehensive set of benchmarks spanning multiple tasks and domains. Experimental results demonstrate that our embedding and reranking models achieve state-of-the-art performance, performing competitively against leading proprietary models in several retrieval tasks. For example, the flagship model Qwen3-8B-Embedding attains a score of 70.58 on the MTEB Multilingual benchmark (Enevoldsen et al., 2025) and 80.68 on the MTEB Code benchmark (Enevoldsen et al., 2025), surpassing the previous state-of-the-art proprietary embedding model, Gemini-Embedding (Lee et al., 2025b). Moreover, our reranking model delivers competitive results across a range of retrieval tasks. The Qwen3-Reranker-0.6B model exceeds previously top-performing models in numerous retrieval tasks, while the larger Qwen3-Reranker-8B model demonstrates even superior performance, improving ranking results by 3.0 points over the 0.6B model across multiple tasks. Furthermore, we include a constructive ablation study to elucidate the key factors contributing to the superior performance of the Qwen3 Embedding series, providing insights into its effectiveness.

In the following sections, we describe the design of the model architecture, detail the training procedures, present the experimental results for both the embedding and reranking models of the Qwen3 Embedding Series, and conclude this technical report by summarizing the key findings and outlining potential directions for future research.

2 Model Architecture

The core idea behind embedding and reranking models is to evaluate relevance in a task-aware manner. Given a query q and a document d, embedding and reranking models assess their relevance based on a similarity criterion defined by instruction I. To enable the models for task-aware relevance estimation, training data is often organized as $\{I_i, q_i, d_i^+, d_{i,1}^-, \cdots, d_{i,n}^-\}$, where d_i^+ represents a

涉及使用伯特（Bert）（例如伯特（Bert））作为培训的基础模型（Reimers＆Gurevych，2019年）的涉及。 LLMS固有的富裕世界知识，文本理解和推理能力导致了对这些体系结构训练的模型的进一步增强。此外，已经进行了大量研究，促进了LLMS整合到诸如训练数据合成和质量数据过滤之类的过程中（Wang等，2024; Lee等，2024; 2025b）。 LLM的基本特征也激发了新的培训范式的引入。例如，在嵌入模型训练过程中，在诸如指导类型，域和语言等方面的差异化任务中纳入了差异化任务，从而提高了下游任务的性能（Su等，2023）。同样，对于重新训练模型培训，通过基于用户提示和结合监督微调的方法来实现进步（Ma等，2023; Pradeep et al。，2023; Zhang et al。，2024a ; Zhuang et al。; Zhuang等，2024）。

在这项工作中，我们介绍了QWEN3嵌入式系列模型，该模型是在QWEN3基础模型之上构建的。 QWEN3基金会同时发布了基础和指示模型版本，我们利用了这些模型的强大多语言文本理解和发电能力，以充分意识到它们在训练嵌入和重新骑行模型中的潜力。为了训练嵌入式模型，我们实施了一条多阶段训练管道，该管道涉及大规模无监督的预训练，然后在高质量的数据集中进行监督的微调。我们还使用各种模型检查点合并模型来增强鲁棒性和一般性。 QWEN3指示模型允许有效合成庞大，高质量，多语言和多任务文本相关性数据集。该合成数据用于最初的无监督训练阶段，而在监督训练的第二阶段中，选择了高质量的小规模数据。对于重读模型，我们以类似的方式采用了两阶段的培训计划，包括高质量的监督微调和模型合并阶段。基于QWEN3骨干模型的不同尺寸（包括0.6B，4B和8B），我们最终训练了三个文本嵌入模型和三个文本重新管理模型。为了促进其在下游任务中的应用，QWEN3嵌入系列系列支持几个实用功能，例如用于嵌入模型的灵活维度表示以及用于嵌入和重新固定模型的可自定义说明。

我们在跨越多个任务和域的一组综合基准中评估了QWEN3嵌入式系列。实验结果表明，我们的嵌入和重读模型达到了最先进的性能，在多个检索任务中对领先的专有模型进行了竞争性的表现。例如，MTEB多语言基准（Enevoldsen等，2025）和MTEB Code Benchmark（Enevoldsen等，2025，2025）的MTEB多语言基准（Enevoldsen等，2025）上的旗舰模型QWEN3-8B插件的得分为70.58，并在MTEB Code Benchmark上获得了70.58。 2025b）。此外，我们的Reranking模型在一系列检索任务中提供竞争成果。 QWEN3-RERANKER-0.6B模型在许多检索任务中超过了先前表现最好的模型，而较大的QWEN3-RERANKER-8B模型也显示出卓越的性能，在多个任务中，在0.6B模型中提高了3.0点的排名。此外，我们还包括一项建设性的消融研究，以阐明有助于QWEN3嵌入系列效果的关键因素，从而提供了对其有效性的见解。

在以下各节中，我们描述了模型体系结构的设计，详细介绍了训练程序，介绍了QWEN3嵌入式系列的嵌入和重新疗法模型的实验结果，并通过汇总了关键发现并概述了未来研究的潜在方向来结束这项技术报告。

2个模型体系结构

嵌入和重读模型背后的核心思想是以任务感知方式评估相关性。给定查询q和一个文档d，嵌入和重读模型根据指令I定义的相似性标准评估其相关性。为了启用任务意识相关估计的模型，训练数据通常被组织为$\{I_i, q_i, d_i^+, d_{i,1}^-, \cdots, \cdots, d_{i,n}^-\}$，其中$d_i^+$代表一个

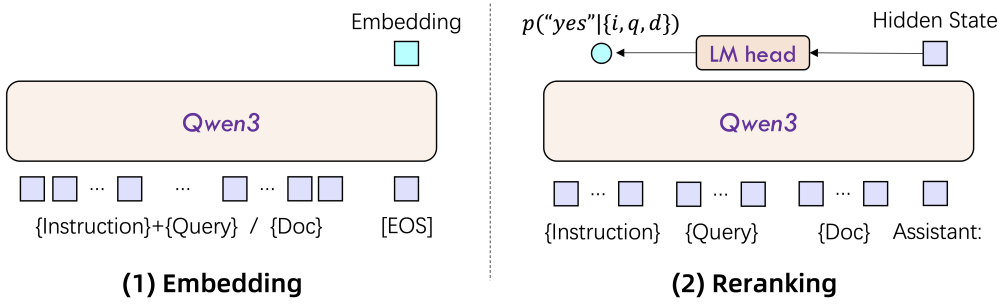

Figure 1: Model architecture of Qwen3-Embedding (left) and Qwen3-Reranker (right).

positive (relevant) document for query q_i, and $d_{i,j}^-$ are negative (irrelevant) documents. Training the model on diverse text pairs broadens its applicability to a range of downstream tasks, including retrieval, semantic textual similarity, classification, and clustering.

Architecture The Qwen3 embedding and reranking models are built on the dense version of Qwen3 foundation models and are available in three sizes: 0.6B, 4B, and 8B parameters. We initialize these models using the Qwen3 foundation models to leverage their capabilities in text modeling and instruction following. The model layers, hidden size, and context length for each model configuration are detailed in Table 1.

Embedding Models For text embeddings, we utilize LLMs with causal attention, appending an `[EOS]` token at the end of the input sequence. The final embedding is derived from the hidden state of the last layer corresponding to this `[EOS]` token.

To ensure embeddings follow instructions during downstream tasks, we concatenate the instruction and the query into a single input context, while leaving the document unchanged before processing with LLMs. The input format for queries is as follows:

```
{Instruction} {Query}<|endoftext|>
```

Reranking Models To more accurately evaluate text similarity, we employ LLMs for point-wise reranking within a single context. Similar to the embedding model, to enable instruction-following capability, we include the instruction in the input context. We use the LLM chat template and frame the similarity assessment task as a binary classification problem. The input to LLMs adheres to the template shown below:

```
<|im_start|>system
Judge whether the Document meets the requirements based on the Query and the
↪  Instruct provided. Note that the answer can only be "yes" or
↪  "no".<|im_end|>
<|im_start|>user
<Instruct>: {Instruction}
<Query>: {Query}
<Document>: {Document}<|im_end|>
<|im_start|>assistant
<think>\n\n</think>\n\n
```

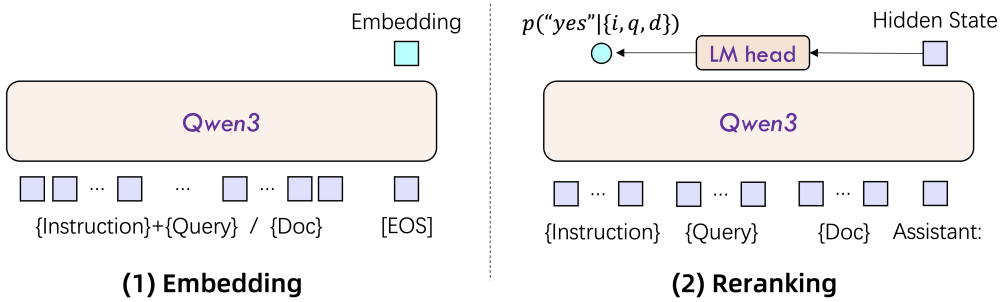

图1：Qwen3-插入（左）和Qwen3速记器（右）的模型结构。

查询q_i和$d_{i,j}^-$的正（相关）文档是负（无关）文档。训练模型对各种文本对，将其适用性扩大到一系列下游任务，包括检索，语义文本相似性，分类和聚类。

体系结构QWEN3嵌入和重新管理模型建立在QWEN3 Foundation Models的密集版本上，可提供三种尺寸：0.6B，4B和8B参数。我们使用QWEN3基础模型初始化这些模型，以利用其在文本建模和说明下的功能。表1详细介绍了每个模型配置的模型层，隐藏大小和上下文长度。

嵌入文本嵌入的模型，我们使用因果关注的LLM，并在输入序列结束时附加了一个[EOS]令牌。最终嵌入是从与此[EOS]令牌相对应的最后一层的隐藏状态得出的。

为了确保嵌入在下游任务期间遵循说明，我们将指令和查询置于单个输入上下文中，同时在使用LLMS处理之前将文档保持不变。查询的输入格式如下：

```
{Instruction} {Query}<|endoftext|>
```

重新计算模型以更准确地评估文本相似性，我们在单个上下文中采用LLM来进行点重新掌握。与嵌入模型类似，为了启用指令跟随功能，我们在输入上下文中包括指令。我们使用LLM聊天模板并将相似性评估任务框架作为二进制分类问题。 LLMS的输入粘附到下面所示的模板：

```
<|im_start|>system
Judge whether the Document meets the requirements based on the Query and the
↪  Instruct provided. Note that the answer can only be "yes" or
↪  "no".<|im_end|>
<|im_start|>user
<Instruct>: {Instruction}
<Query>: {Query}
<Document>: {Document}<|im_end|>
<|im_start|>assistant
<think>\n\n</think>\n\n
```

Model Type	Models	Size	Layers	Sequence Length	Embedding Dimension	MRL Support	Instruction Aware
Text Embedding	Qwen3-Embedding-0.6B	0.6B	28	32K	1024	Yes	Yes
	Qwen3-Embedding-4B	4B	36	32K	2560	Yes	Yes
	Qwen3-Embedding-8B	8B	36	32K	4096	Yes	Yes
Text Reranking	Qwen3-Reranker-0.6B	0.6B	28	32K	-	-	Yes
	Qwen3-Reranker-4B	4B	36	32K	-	-	Yes
	Qwen3-Reranker-8B	8B	36	32K	-	-	Yes

Table 1: Model architecture of Qwen3 Embedding models. "MRL Support" indicates whether the embedding model supports custom dimensions for the final embedding. "Instruction Aware" notes whether the embedding or reranker model supports customizing the input instruction according to different tasks.

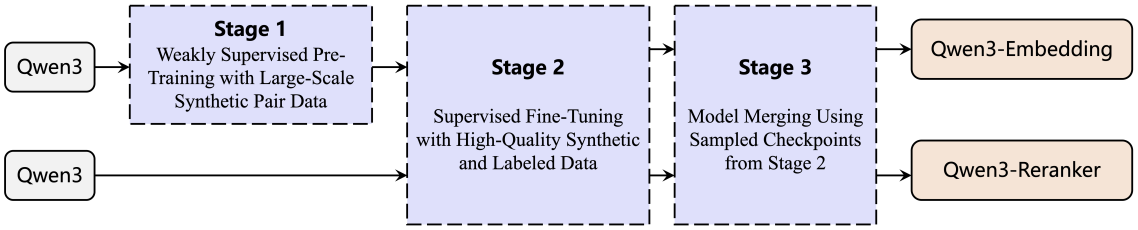

Figure 2: Training pipeline of Qwen3 Embedding and Reranking models.

To calculate the relevance score based on the given input, we assess the likelihood of the next token being "yes" or "no." This is expressed mathematically as follows:

$$\text{score}(q,d) = \frac{e^{P(\text{yes}|I,q,d)}}{e^{P(\text{yes}|I,q,d)} + e^{P(\text{no}|I,q,d)}}$$

3 Models Training

In this section, we describe the multi-stage training pipeline adopted and present the key elements of this training recipe, including training objective, training data synthesis, and filtering of high-quality training data.

3.1 Training Objective

Before introducing our training pipeline, we first outline the optimized loss functions used for the embedding and reranking models during the training process. For the embedding model, we utilize an improved contrastive loss based on the InfoNCE framework (Oord et al., 2018). Given a batch of N training instances, the loss is defined as:

$$L_{\text{embedding}} = -\frac{1}{N}\sum_{i}^{N} \log \frac{e^{(s(q_i,d_i^+)/\tau)}}{Z_i}, \tag{1}$$

where $s(\cdot,\cdot)$ is a similarity function (we use cosine similarity), τ is a temperature parameter, and Z_i is the normalization factor that aggregates the similarity scores of the positive pair against various negative pairs:

$$Z_i = e^{(s(q_i,d_i^+)/\tau)} + \sum_{k}^{K} m_{ik} e^{(s(q_i,d_{i,k}^-)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(q_i,q_j)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(d_i^+,d_j)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(q_i,d_j)/\tau)}$$

Model Type	Models	Size	Layers	Sequence Length	Embedding Dimension	MRL Support	Instruction Aware
Text Embedding	Qwen3-Embedding-0.6B	0.6B	28	32K	1024	Yes	Yes
	Qwen3-Embedding-4B	4B	36	32K	2560	Yes	Yes
	Qwen3-Embedding-8B	8B	36	32K	4096	Yes	Yes
Text Reranking	Qwen3-Reranker-0.6B	0.6B	28	32K	-	-	Yes
	Qwen3-Reranker-4B	4B	36	32K	-	-	Yes
	Qwen3-Reranker-8B	8B	36	32K	-	-	Yes

表1：QWEN3嵌入模型的模型架构。 “ MRL支持”表示嵌入模型是否支持最终嵌入的自定义维度。 “指导意识到”指出，嵌入式或重读者模型是否支持根据不同任务自定义输入指令。

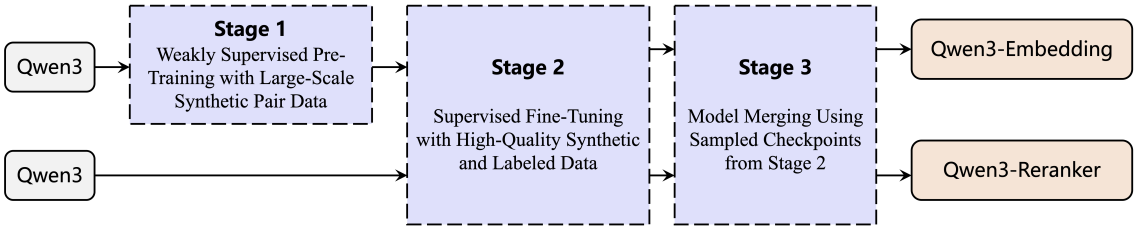

图2：QWEN3嵌入和重新管理模型的训练管道。

为了根据给定的输入计算相关性得分，我们评估了下一个令牌为“是”或“否”的可能性。这在数学上表示如下：

$$\text{score}(q,d) = \frac{e^{P(\text{yes}|I,q,d)}}{e^{P(\text{yes}|I,q,d)} + e^{P(\text{no}|I,q,d)}}$$

3个型号培训

在本节中，我们描述了采用的多阶段培训管道，并介绍了该培训配方的关键要素，包括培训目标，培训数据综合和过滤高质量的培训数据。

3.1培训目标

在介绍培训管道之前，我们首先概述了在培训过程中用于嵌入和重新固定模型的优化损失功能。对于嵌入模型，我们根据Infonce框架利用改进的对比度损失（Oord等，2018）。给定一批N培训实例，损失定义为：

$$L_{\text{embedding}} = -\frac{1}{N}\sum_{i}^{N}\log\frac{e^{(s(q_i,d_i^+)/\tau)}}{Z_i}, \tag{1}$$

其中$s(\cdot,\cdot)$是一个相似性函数（我们使用余弦相似性），τ是一个温度参数，Z_i是汇总的归一化因子，它汇总了与各种负面对的阳性对的相似性得分：

$$Z_i = e^{(s(q_i,d_i^+)/\tau)} + \sum_{k}^{K} m_{ik} e^{(s(q_i,d_{i,k}^-)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(q_i,q_j)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(d_i^+,d_j)/\tau)} + \sum_{j\neq i} m_{ij} e^{(s(q_i,d_j)/\tau)}$$

where these terms represent similarities with: (1) the positive document d_i^+, (2) K hard negatives $d_{i,k}^-$, (3) other in-batch queries q_j, (4) other in-batch documents d_j compared against the positive document d_i^+. (5) other in-batch documents d_j compared against the query q_i. The mask factor m_{ij} is designed to mitigate the impact of false negatives and is defined as:

$$m_{ij} = \begin{cases} 0 & \text{if } s_{ij} > s(q_i, d_i^+) + 0.1 \text{ or } d_j == d_i^+, \\ 1 & \text{otherwise,} \end{cases}$$

among which s_{ij} is the corresponding score of q_i, d_j or q_i, q_j.

For the reranking model, we optimize the Supervised Fine-Tuning (SFT) loss defined as:

$$L_{\text{reranking}} = -\log p(l|\mathcal{P}(q,d)), \tag{2}$$

where $p(\cdot|*)$ denotes the probability assigned by LLM. The label l is "yes" for positive documents and "no" for negatives. This loss function encourages the model to assign higher probabilities to correct labels, thereby improving the ranking performance.

3.2 Multi-stage Training

The multi-stage training approach is a common practice for training text embedding models (Li et al., 2023; Wang et al., 2022; Chen et al., 2024). This strategy typically begins with initial training on large-scale, semi-supervised data that includes noise, followed by fine-tuning using smaller, high-quality supervised datasets. This two-step process enhances the performance and generalization capabilities of embedding models. Large-scale weakly supervised training data contribute significantly to the model's generalization, while fine-tuning with high-quality data in subsequent stages further improves model performance. Both stages of training for embedding models utilize the optimization objective defined in Equation 1, whereas the reranking model training employs the loss function defined in Equation 2 as the optimization target.

Building upon the existing multi-stage training framework, the Qwen3 Embedding series introduces the following key innovations:

- Large-Scale Synthetic Data-Driven Weak Supervision Training: Unlike previous works (e.g., GTE, E5, BGE models), where weakly supervised training data are primarily collected from open-source communities such as Q&A forums or academic papers, we propose leveraging the text understanding and generation capabilities of foundation models to synthesize pair data directly. This approach allows for arbitrary definition of various dimensions of the desired pair data, such as task, language, length, and difficulty within the synthesis prompts. Compared to data collection from open-domain sources, foundation model-driven data synthesis offers greater controllability, enabling precise management of the quality and diversity of the generated data, particularly in low-resource scenarios and languages.
- High-Quality Synthetic Data Utilization in Supervised Fine Tuning: Due to the exceptional performance of the Qwen3 Foundation model, the synthesized data is of notably high quality. Therefore, in the second stage of supervised training, selective incorporation of this high-quality synthetic data further enhances the overall model performance and generalization capabilities.
- Model Merging: Inspired by previous work (Li et al., 2024), after completing the supervised fine-tuning, we applied a model merging technique based on spherical linear interpolation (slerp). This technique involves merging multiple model checkpoints saved during the fine-tuning process. This step aims to boost the model's robustness and generalization performance across various data distributions.

It is important to note that the reranking model's training process does not include a first-stage weakly supervised training phase.

这些术语表示与以下方式相似之处：（1）正面文档d_i^+，（2）K硬负$d_{i,k}^-$，（3）与正面文档d_i^+相比。（5）与查询q_i相比，其他批处理文档d_j。掩模因子m_{ij}旨在减轻虚假负面的影响，并定义为：

$$m_{ij} = \begin{cases} 0 & \text{if } s_{ij} > s(q_i, d_i^+) + 0.1 \text{ or } d_j == d_i^+, \\ 1 & \text{otherwise,} \end{cases}$$

其中s_{ij}是q_i，d_j或q_i，q_j的相应分数。

对于重新脉模式 l，我们优化了监督的微调（SFT）损失defi ned as：

$$L_{\text{reranking}} = -\log p(l|\mathcal{P}(q,d)), \tag{2}$$

其中$p(\cdot|*)$表示LLM分配的概率。对于积极文档而言，标签l是“是”，而对负面文档则是“否”。此损失功能鼓励模型分配更高的概率以纠正标签，从而提高排名性能。

3.2多阶段培训

多阶段训练方法是训练文本嵌入模型的常见实践（Li等，2023； Wang等，2022； Chen等，2024）。该策略通常始于大规模，半监督数据的初步培训，其中包括噪声，然后使用较小的高质量监督数据集进行微调。这个两步过程增强了嵌入模型的性能和概括能力。大规模弱监督的培训数据对模型的概括产生了重大贡献，同时在随后的阶段进行高质量数据进行微调进一步改善了模型性能。嵌入模型的训练的两个阶段都利用了等式1中定义的优化目标，而重读模型培训则采用公式2中定义的损耗函数作为优化目标。

在现有的多阶段培训框架的基础上，QWEN3嵌入式系列介绍了以下关键创新：

- 大规模合成数据驱动的弱监督培训：与以前的作品（例如GTE，E5，BGE模型）不同，弱监督的培训数据主要是从开源社区中收集的，例如问答论坛或学术报纸，例如，我们提议利用基础模型的文本理解和生成能力，以直接合成配对数据。这种方法允许对所需对数据的各个维度进行任意定义，例如任务，语言，长度和困难，并在综合提示中。与开放域源的数据收集相比，基础模型驱动的数据合成提供了更大的可控性，从而可以精确地管理生成数据的质量和多样性，尤其是在低资源的场景和语言中。

- 高质量的合成数据在监督的微调中使用：由于QWEN3基础模型的出色性能，合成的数据的质量尤其高。因此，在监督培训的第二阶段，这种高质量合成数据的选择性合并进一步增强了整体模型性能和泛化能力。

- 模型合并：受到先前工作的启发（Li等，2024），在完成监督的微调后，我们应用了基于球形线性插值（SLERP）的模型合并技术。该技术涉及合并在微调过程中保存的多个模型检查点。此步骤旨在提高模型在各种数据分布中的稳健性和泛化性能。

重要的是要注意，Reranking模型的培训过程不包括第一阶段弱监督的培训阶段。

3.3 Synthetic Dataset

To create a robust synthetic dataset for training models on various similarity tasks, we generate diverse text pairs spanning categories such as retrieval, bitext mining, classification, and semantic textual similarity (STS). The quality of these synthetic data pairs is ensured by utilizing the Qwen3-32B model as the foundational model for data synthesis. We have designed a diverse prompting strategy to improve the variety and authenticity of the generated data. For instance, in the text retrieval task, we synthesize data using the multilingual pre-training corpus from Qwen3. During the data synthesis process, specific roles are assigned to each document to simulate potential users querying that document. This injection of user perspectives enhances the diversity and realism of the synthetic queries. Specifically, we utilize a retrieval model to identify the top five role candidates for each document from a role library and present these documents along with their role candidates to the prompt. This guides the model in outputting the most suitable role configuration for query generation. Moreover, the prompt incorporates various dimensions such as query type (e.g., keyword, factual, summary, judgment), query length, difficulty, and language. This multidimensional approach ensures the quality and diversity of the synthetic data.

Finally, we create a total of approximately 150 million pairs of multi-task weak supervision training data. Our experiments reveal that the embedding model trained with these synthetic data performs exceptionally well in downstream evaluations, particularly surpassing many previously supervised models in the MTEB Multilingual benchmarks. This motivates us to filter the synthetic data to identify high-quality pairs for inclusion in a second stage of supervised training. We employ a simple cosine similarity calculation to select data pairs, retaining those with a cosine similarity greater than 0.7 from randomly sampled data. Ultimately, approximately 12 million high-quality supervised training data pairs are selected for further training.

Model	**Size**	**Mean (Task)**	**Mean (Type)**	Bitext Mining	Class-ification	Clus-tering	Inst. Retrieval	Multilabel Class.	Pair Class.	Rerank	Retrieval	STS
Selected Open-Source Models												
NV-Embed-v2	7B	56.29	49.58	57.84	57.29	40.80	1.04	18.63	78.94	63.82	56.72	71.10
GritLM-7B	7B	60.92	53.74	70.53	61.83	49.75	3.45	22.77	79.94	63.78	58.31	73.33
BGE-M3	0.6B	59.56	52.18	79.11	60.35	40.88	-3.11	20.1	80.76	62.79	54.60	74.12
multilingual-e5-large-instruct	0.6B	63.22	55.08	80.13	64.94	50.75	-0.40	22.91	80.86	62.61	57.12	76.81
gte-Qwen2-1.5B-instruct	1.5B	59.45	52.69	62.51	58.32	52.05	0.74	24.02	81.58	62.58	60.78	71.61
gte-Qwen2-7b-Instruct	7B	62.51	55.93	73.92	61.55	52.77	4.94	25.48	85.13	65.55	60.08	73.98
Commercial APIs												
text-embedding-3-large	-	58.93	51.41	62.17	60.27	46.89	-2.68	22.03	79.17	63.89	59.27	71.68
Cohere-embed-multilingual-v3.0	-	61.12	53.23	70.50	62.95	46.89	-1.89	22.74	79.88	64.07	59.16	74.80
Gemini Embedding	-	68.37	59.59	79.28	71.82	54.59	5.18	**29.16**	83.63	65.58	67.71	79.40
Qwen3 Embedding Models												
Qwen3-Embedding-0.6B	0.6B	64.33	56.00	72.22	66.83	52.33	5.09	24.59	80.83	61.41	64.64	76.17
Qwen3-Embedding-4B	4B	69.45	60.86	79.36	72.33	57.15	**11.56**	26.77	85.05	65.08	69.60	80.86
Qwen3-Embedding-8B	8B	**70.58**	**61.69**	**80.89**	**74.00**	**57.65**	10.06	28.66	**86.40**	**65.63**	**70.88**	**81.08**

Table 2: Performance on MTEB Multilingual (Enevoldsen et al., 2025). For compared models, the scores are retrieved from MTEB online leaderboard on June 4th, 2025.

4 Evaluation

We conduct comprehensive and fair evaluations across multiple benchmarks to assess the capabilities of Qwen3 Embedding models.

4.1 Settings

For the text embedding models, we utilize the Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025) for evaluation. MMTEB is a large-scale, community-driven expansion of MTEB (Muennighoff et al., 2023), covering over 500 quality-controlled evaluation tasks

3.3合成数据集

为了创建一个可在各种相似性任务的培训模型的强大合成数据集，我们生成跨越类别的各种文本对，例如检索，bitext挖掘，分类和语义文本相似性（STS）。这些合成数据对的质量可以通过利用QWEN3-32B模型作为数据合成的基础模型来确保。我们设计了一种多样化的提示策略，以改善生成数据的多样性和真实性。例如，在文本检索任务中，我们使用QWEN3的多语言预训练语料库合成数据。在数据综合过程中，将特定角色分配给每个文档，以模拟潜在用户查询该文档。这种注入用户观点增强了合成查询的多样性和现实性。具体来说，我们利用检索模型从角色库中确定每个文档的前五名候选者，并将这些文档及其角色候选人介绍给提示。这将指导模型输出最合适的查询生成角色配置。此外，提示包括各种维度，例如查询类型（例如关键字，事实，摘要，判断），查询长度，难度和语言。这种多维方法可确保合成数据的质量和多样性。

最后，我们创建了约1.5亿对多任务弱监督培训数据。我们的实验表明，使用这些合成数据训练的嵌入模型在下游评估中表现出色，尤其是超过MTEB多语言基准中的许多先前监督模型。这促使我们过滤综合数据，以识别高质量对以在监督训练的第二阶段中纳入。我们采用一个简单的余弦相似性计算来选择数据对，从而从随机采样数据中保留大于0.7的余弦相似性的数据对。最终，选择了大约1200万个高质量的监督培训数据对进行进一步培训。

Model	Size	Mean (Task)	Mean (Type)	Bitext Mining	Class-ification	Clus-tering	Inst. Retrieval	Multilabel Class.	Pair Class.	Rerank	Retrieval	STS
Selected Open-Source Models												
NV-Embed-v2	7B	56.29	49.58	57.84	57.29	40.80	1.04	18.63	78.94	63.82	56.72	71.10
GritLM-7B	7B	60.92	53.74	70.53	61.83	49.75	3.45	22.77	79.94	63.78	58.31	73.33
BGE-M3	0.6B	59.56	52.18	79.11	60.35	40.88	-3.11	20.1	80.76	62.79	54.60	74.12
multilingual-e5-large-instruct	0.6B	63.22	55.08	80.13	64.94	50.75	-0.40	22.91	80.86	62.61	57.12	76.81
gte-Qwen2-1.5B-instruct	1.5B	59.45	52.69	62.51	58.32	52.05	0.74	24.02	81.58	62.58	60.78	71.61
gte-Qwen2-7b-Instruct	7B	62.51	55.93	73.92	61.55	52.77	4.94	25.48	85.13	65.55	60.08	73.98
Commercial APIs												
text-embedding-3-large	-	58.93	51.41	62.17	60.27	46.89	-2.68	22.03	79.17	63.89	59.27	71.68
Cohere-embed-multilingual-v3.0	-	61.12	53.23	70.50	62.95	46.89	-1.89	22.74	79.88	64.07	59.16	74.80
Gemini Embedding	-	68.37	59.59	79.28	71.82	54.59	5.18	**29.16**	83.63	65.58	67.71	79.40
Qwen3 Embedding Models												
Qwen3-Embedding-0.6B	0.6B	64.33	56.00	72.22	66.83	52.33	5.09	24.59	80.83	61.41	64.64	76.17
Qwen3-Embedding-4B	4B	69.45	60.86	79.36	72.33	57.15	**11.56**	26.77	85.05	65.08	69.60	80.86
Qwen3-Embedding-8B	8B	**70.58**	**61.69**	**80.89**	**74.00**	**57.65**	10.06	28.66	**86.40**	**65.63**	**70.88**	**81.08**

表2：MTEB多语言上的性能（Enevoldsen等，2025）。对于比较的模型，分数将于2025年6月4日从MTEB Online排行榜上检索。

4评估

我们对多个基准进行全面和公平的评估，以评估QWEN3嵌入模型的功能。

4.1设置

对于文本嵌入模型，我们利用大量的多语言文本嵌入基准（MMTEB）（Enevoldsen等，2025）进行评估。 MMTEB是MTEB的大规模，社区驱动的扩展（Muennighoff等，2023），涵盖了500多个质量控制的评估任务

Model	Size	Dim	MTEB (Eng, v2)		CMTEB		MTEB (Code)
			Mean (Task)	Mean (Type)	Mean (Task)	Mean (Type)	
Selected Open-Source Models							
NV-Embed-v2	7B	4096	69.81	65.00	63.0	62.0	-
GritLM-7B	7B	4096	67.07	63.22	-	-	73.6[α]
multilingual-e5-large-instruct	0.6B	1024	65.53	61.21	-	-	65.0[α]
gte-Qwen2-1.5b-instruct	1.5B	1536	67.20	63.26	67.12	67.79	-
gte-Qwen2-7b-instruct	7B	3584	70.72	65.77	71.62	72.19	56.41[γ]
Commercial APIs							
text-embedding-3-large	-	3072	66.43	62.15	-	-	58.95[γ]
cohere-embed-multilingual-v3.0	-	1024	66.01	61.43	-	-	51.94[γ]
Gemini Embedding	-	3072	73.30	67.67	-	-	74.66[γ]
Qwen3 Embedding Models							
Qwen3-Embedding-0.6B	0.6B	1024	70.70	64.88	66.33	67.44	75.41
Qwen3-Embedding-4B	4B	2560	74.60	68.09	72.26	73.50	80.06
Qwen3-Embedding-8B	8B	4096	**75.22**	**68.70**	**73.83**	**75.00**	**80.68**

Table 3: Performance on MTEB Engilish, MTEB Chinese, MTEB Code. [α]Taken from (Enevoldsen et al., 2025). [γ]Taken from (Lee et al., 2025b). For other compared models, the scores are retrieved from MTEB online leaderboard on June 4th, 2025.

across more than 250 languages. In addition to classic text tasks such as as a variety of retrieval, classification, and semantic textual similarity, MMTEB includes a diverse set of challenging and novel tasks, such as instruction following, long-document retrieval, and code retrieval, representing the largest multilingual collection of evaluation tasks for embedding models to date. Our MMTEB evaluations encompass 216 individual evaluation tasks, consisting of 131 tasks for MTEB (Multilingual) (Enevoldsen et al., 2025), 41 tasks for MTEB (English, v2) (Muennighoff et al., 2023), 32 tasks for CMTEB (Xiao et al., 2024), and 12 code retrieval tasks for MTEB (Code) (Enevoldsen et al., 2025).

Moreover, we select a series of text retrieval tasks to assess the text reranking capabilities of our models. We explore three types of retrieval tasks: (1) Basic Relevance Retrieval, categorized into English, Chinese, and Multilingual, evaluated on MTEB (Muennighoff et al., 2023), CMTEB (Xiao et al., 2024), MMTEB (Enevoldsen et al., 2025), and MLDR (Chen et al., 2024), respectively; (2) Code Retrieval, evaluated on MTEB-Code (Enevoldsen et al., 2025), which comprises only code-related retrieval data.; and (3) Complex Instruction Retrieval, evaluated on FollowIR (Weller et al., 2024).

Compared Methods We compare our models with the most prominent open-source text embedding models and commercial API services. The open-source models include the GTE (Li et al., 2023; Zhang et al., 2024b), E5 (Wang et al., 2022), and BGE (Xiao et al., 2024) series, as well as NV-Embed-v2 (Lee et al., 2025a), GritLM-7B Muennighoff et al. (2025). The commercial APIs evaluated are text-embedding-3-large from OpenAI, Gemini-embedding from Google, and Cohere-embed-multilingual-v3.0. For reranking, we compare with the rerankers of jina[1], mGTE (Zhang et al., 2024b) and BGE-m3 (Chen et al., 2024).

4.2 Main Results

Embedding In Table 2, we present the evaluation results on MMTEB (Enevoldsen et al., 2025), which comprehensively covers a wide range of embedding tasks across multiple languages. Our Qwen3-Embedding-4B/8B models achieve the best performance, and our smallest model, Qwen3-Embedding-0.6B, only lags behind the best-performing baseline method (Gemini-Embedding), despite having only 0.6B parameters. In Table 3, we present the evaluation results on MTEB (English, v2) (Muennighoff et al., 2023), CMTEB (Xiao et al., 2024), and MTEB (Code) (Enevoldsen et al., 2025). The scores reflect similar trends as MMTEB, with our Qwen3-Embedding-4B/8B models

[1] https://hf.co/jinaai/jina-reranker-v2-base-multilingual

Model	Size	Dim	MTEB (Eng, v2)		CMTEB		MTEB (Code)
			Mean (Task)	Mean (Type)	Mean (Task)	Mean (Type)	
Selected Open-Source Models							
NV-Embed-v2	7B	4096	69.81	65.00	63.0	62.0	-
GritLM-7B	7B	4096	67.07	63.22	-	-	73.6[α]
multilingual-e5-large-instruct	0.6B	1024	65.53	61.21	-	-	65.0[α]
gte-Qwen2-1.5b-instruct	1.5B	1536	67.20	63.26	67.12	67.79	-
gte-Qwen2-7b-instruct	7B	3584	70.72	65.77	71.62	72.19	56.41[γ]
Commercial APIs							
text-embedding-3-large	-	3072	66.43	62.15	-	-	58.95[γ]
cohere-embed-multilingual-v3.0	-	1024	66.01	61.43	-	-	51.94[γ]
Gemini Embedding	-	3072	73.30	67.67	-	-	74.66[γ]
Qwen3 Embedding Models							
Qwen3-Embedding-0.6B	0.6B	1024	70.70	64.88	66.33	67.44	75.41
Qwen3-Embedding-4B	4B	2560	74.60	68.09	72.26	73.50	80.06
Qwen3-Embedding-8B	8B	4096	**75.22**	**68.70**	**73.83**	**75.00**	**80.68**

表3：MTEB Engilish，MTEB中文，MTEB代码的性能。[α]取自（Enevoldsen等，2025）。[γ]取自（Lee等，2025b）。对于其他比较模型，分数将于2025年6月4日从MTEB Online排行榜中获取。

遍及250多种语言。除了经典的文本任务（例如各种检索，分类和语义文本相似性）外，MMTEB还包括各种具有挑战性和新颖的任务，例如以下教学，长期记录检索和代码检索，代表了最大的多项式评估任务收集到日期的嵌入方式。我们的MMTEB评估包括216个个人评估任务，由131个MTEB（Multilin-gual）任务组成（Enevoldsen等，2025），41个MTEB（英语，V2）任务（Muennighoff，V2）（Muennighoff et al。，2023），32个CMTEB（CMTEB（CMTEB）的任务，cmteb（xiao等人）（Xiao等人，2024）。（代码）（Enevoldsen等，2025）。

此外，我们选择了一系列文本检索任务来评估模型的文本重新依据功能。我们探讨了三种类型的检索任务：（1）在MTEB上评估的基本相关性检索，分为英语，中文和多语言，并在MTEB上进行评估（Muennighoff等，2023），CMTEB（Xiao等，2024），MMTEB，MMTEB，MMTEB（Enevoldsen等，2025），和Mld（2025），和2024;（2）代码检索，对MTEB代码进行了评估（Enevoldsen等，2025），该代码仅包含与代码相关的检索数据。（3）复杂的指令检索，在Collowir上进行了评估（Weller等，2024）。

比较方法，我们将模型与最突出的开源文本嵌入模型和商业API服务进行了比较。开源模型包括GTE（Li等，2023; Zhang et al。，2024b），E5（Wang等，2022）和BGE（Xiao等，2024）系列，以及NV-Embed-V2（Lee等，2025A），Gritlm-7b Muennegr。（2025）。评估的商业API是来自Openai的Text-ebbedding-3-Large，Google的双子座装饰，以及Cohere-Exbed-多语言-V3.0。为了重读，我们将其与Jina [1]，Mgte（Zhang等，2024b）和BGE-M3（Chen等，2024）的Reranker进行比较。

4.2主要结果

嵌入表2中，我们在MMTEB上介绍了评估结果（Enevoldsen等，2025），该结果全面涵盖了多种语言的各种嵌入任务。我们的QWEN3-插入-4B/8B型号实现了最佳性能，并且我们最小的型号QWEN3-嵌入式-0.6B，尽管只有0.6B参数，但仅落后于表现最佳的基线方法（Gemini-Embedding）。在表3中，我们列出了MTEB（英语，V2）（Muennighoff等，2023），CMTEB（Xiao等，2024）和MTEB（代码）（Enevoldsen等，2025）。分数反映了与MMTEB相似的趋势，我们的QWEN3-EBEDDING-4B/8B型号

[1] https://hf.co/jinaai/jina-reranker-v2-base-multilingual

Model	Param	Basic Relevance Retrieval					
		MTEB-R	CMTEB-R	MMTEB-R	MLDR	MTEB-Code	FollowIR
Qwen3-Embedding-0.6B	0.6B	61.82	71.02	64.64	50.26	75.41	5.09
Jina-multilingual-reranker-v2-base	0.3B	58.22	63.37	63.73	39.66	58.98	-0.68
gte-multilingual-reranker-base	0.3B	59.51	74.08	59.44	66.33	54.18	-1.64
BGE-reranker-v2-m3	0.6B	57.03	72.16	58.36	59.51	41.38	-0.01
Qwen3-Reranker-0.6B	0.6B	65.80	71.31	66.36	67.28	73.42	5.41
Qwen3-Reranker-4B	4B	**69.76**	75.94	72.74	69.97	81.20	**14.84**
Qwen3-Reranker-8B	8B	69.02	**77.45**	**72.94**	**70.19**	**81.22**	8.05

Table 4: Evaluation results for reranking models. We use the retrieval subsets of MTEB(eng, v2), MTEB(cmn, v1) and MMTEB, which are MTEB-R, CMTEB-R and MMTEM-R. The rest are all retrieval tasks. All scores are our runs based on the retrieval top-100 results from the first row.

Model	MMTEB	MTEB (Eng, v2)	CMTEB	MTEB (Code, v1)
Qwen3-Embedding-0.6B w/ only synthetic data	58.49	60.63	59.78	66.79
Qwen3-Embedding-0.6B w/o synthetic data	61.21	65.59	63.37	74.58
Qwen3-Embedding-0.6B w/o model merge	62.56	68.18	64.76	74.89
Qwen3-Embedding-0.6B	64.33	70.70	66.33	75.41

Table 5: Performance (mean task) on MMTEB, MTEB(eng, v2), CMTEB and MTEB(code, v1) for Qwen3-Embedding-0.6B model with different training setting.

consistently outperforming others. Notably, the Qwen3-Embedding-0.6B model ranks just behind the Gemini-Embedding, while being competitive with the gte-Qwen2-7B-instruct.

Reranking In Table 4, we present the evaluation results on various reranking tasks (§4.1). We utilize the Qwen3-Embedding-0.6B model to retrieve the top-100 candidates and then apply different reranking models for further refinement. This approach ensures a fair evaluation of the reranking models. Our results indicate that all three Qwen3-Reranker models enhance performance compared to the embedding model and surpass all baseline reranking methods, with Qwen3-Reranker-8B achieving the highest performance across most tasks.

4.3 Analysis

To further analyze and explore the key elements of the Qwen3 Embedding model training framework, we conduct an analysis from the following dimensions:

Effectiveness of Large-Scale Weakly Supervised Pre-Training We first analyze the effectiveness of the large-scale weak supervised training stage for the embedding models. As shown in Table 5, the Qwen3-Embedding-0.6B model trained solely on synthetic data (without subsequent training stages, as indicated in the first row) achieves reasonable and strong performance compared to the final Qwen3-Embedding-0.6B model (as shown in the last row). If we further remove the weak supervised training stage (i.e., without synthetic data training, as seen in the second row), the final performance shows a clear decline. This indicates that the large-scale weak supervised training stage is crucial for achieving superior performance.

Effectiveness of Model Merging Next, we compare the performance differences arising from the model merging stage. As shown in Table 5, the model trained without model merging techniques (the third row, which uses data sampling to balance various tasks) performs considerably worse than the final Qwen3-Embedding-0.6B model (which employs model merging, as shown in the last row). This indicates that the model merging stage is also critical for developing strong models.

Model	Param	Basic Relevance Retrieval					
		MTEB-R	CMTEB-R	MMTEB-R	MLDR	MTEB-Code	FollowIR
Qwen3-Embedding-0.6B	0.6B	61.82	71.02	64.64	50.26	75.41	5.09
Jina-multilingual-reranker-v2-base	0.3B	58.22	63.37	63.73	39.66	58.98	-0.68
gte-multilingual-reranker-base	0.3B	59.51	74.08	59.44	66.33	54.18	-1.64
BGE-reranker-v2-m3	0.6B	57.03	72.16	58.36	59.51	41.38	-0.01
Qwen3-Reranker-0.6B	0.6B	65.80	71.31	66.36	67.28	73.42	5.41
Qwen3-Reranker-4B	4B	**69.76**	75.94	72.74	69.97	81.20	**14.84**
Qwen3-Reranker-8B	8B	69.02	**77.45**	**72.94**	**70.19**	**81.22**	8.05

表4：Reranking模型的评估结果。我们使用MTEB（ENG，V2），MTEB（CMN，V1）和MMTEB的检索子集，即MTEB-R，CMTEB-R和MMTEM-R。其余的都是检索任务。所有分数都是根据第一行的检索前100个结果的运行。

Model	MMTEB	MTEB (Eng, v2)	CMTEB	MTEB (Code, v1)
Qwen3-Embedding-0.6B w/ only synthetic data	58.49	60.63	59.78	66.79
Qwen3-Embedding-0.6B w/o synthetic data	61.21	65.59	63.37	74.58
Qwen3-Embedding-0.6B w/o model merge	62.56	68.18	64.76	74.89
Qwen3-Embedding-0.6B	64.33	70.70	66.33	75.41

表5：MMTEB，MTEB（ENG，V2），CMTEB和MTEB（代码，V1）上的性能（平均任务），用于具有不同训练设置的QWEN3-EMBEDDING-0.6B模型。

始终超过他人。值得注意的是，Qwen3-Embedding-0.6B型号仅落后于Gemini-Exbing，同时与GTE-QWEN2-7B-INSTRUCT具有竞争力。

在表4中，我们介绍了各种重新依据任务的评估结果（第4.1节）。我们利用QWEN3-EMBEDDING-0.6B型号来检索前100名候选者，然后应用不同的重新骑行模型进行进一步的改进。这种方法可确保对重新管理模型的公平评估。我们的结果表明，与嵌入式模型相比，所有三个QWEN3-速航模型都提高了性能，并且超过了所有基线重读方法，QWEN3-Reranker-8B在大多数任务中都达到了最高的性能。

4.3分析

为了进一步分析和探索QWEN3嵌入模型培训框架的关键要素，我们从以下维度进行了分析：

大规模监督的预训练的有效性首先分析了嵌入模型的大规模监督训练阶段的有效性。如表5所示，与最终的QWEN3-EMBEDDING-0.6B模型相比，仅根据合成数据进行训练的QWEN3-EMBEDDING-0.6B模型（如第一行所示，没有随后的训练阶段，如前一行所示）实现了合理且强劲的性能。如果我们进一步删除弱监督训练阶段（即，如第二行所示，没有合成数据训练），则最终表现会明显下降。这表明大规模的监督训练阶段对于实现卓越的表现至关重要。

模型合并的有效性接下来，我们比较模型合并阶段引起的性能差异。如表5所示，未经模型合并技术训练的模型（第三行，使用数据采样来平衡各种任务）的性能比最终的QWEN3-Embedding-0.6B模型（采用模型合并，如上一行所示）。这表明模型合并阶段对于开发强模型也至关重要。

5 Conclusion

In this technical report, we present the Qwen3-Embedding series, a comprehensive suite of text embedding and reranking models based on the Qwen3 foundation models. These models are designed to excel in a wide range of text embedding and reranking tasks, including multilingual retrieval, code retrieval, and complex instruction following. The Qwen3-Embedding models are built upon a robust multi-stage training pipeline that combines large-scale weakly supervised pre-training on synthetic data with supervised fine-tuning and model merging on high-quality datasets. The Qwen3 LLMs play a crucial role in synthesizing diverse training data across multiple languages and tasks, thereby enhancing the models' capabilities. Our comprehensive evaluations demonstrate that the Qwen3-Embedding models achieve state-of-the-art performance across various benchmarks, including MTEB, CMTEB, MMTEB, and several retrieval benchmarks. We are pleased to open-source the Qwen3-Embedding and Qwen3-Reranker models (0.6B, 4B, and 8B), making them available for the community to use and build upon.

References

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.137/.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zl3pfz4VCV.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2553–2561, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=lgsyLSsDRe.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025b.

5结论

在这份技术报告中，我们介绍了QWEN3插入系列，这是一套基于QWEN3基础模型的文本嵌入和重新骑行模型。这些模型旨在在广泛的文本嵌入和重新计算任务中表现出色，包括多语言检索，代码检索和复杂的说明。 QWEN3插入模型建立在强大的多阶段训练管道上，该管道结合了大规模监督的综合数据预训练，并在高质量数据集中合并的模型。 QWEN3 LLM在跨多种语言和任务综合多种培训数据中起着至关重要的作用，从而增强了模型的功能。我们的全面评估表明，QWEN3插入模型在包括MTEB，CMTEB，MMTEB和几个检索基准在内的各种基准的最新性能。我们很高兴为QWEN3插件和QWEN3速记机型号（0.6B，4B和8B）开放源，使它们可以供社区使用和建立。

参考

Jianlyu Chen，Shitao Xiao，Peitian Zhang，Kun Luo，Defu Lian和Zheng Liu。 M3插入：通过自我知识蒸馏，多语言，多功能性，多晶格文本嵌入。在
Findings of the Association for Computational Linguistics: ACL 2024，第2318–2335页，曼谷，泰国，2024年8月。计算语言学协会。 URL https:// aclanthology.org/2024.findings-acl.137/。 Kenneth Enevoldsen，Isaac Chung，Imene Kerboua，M ´Arton Kardos，Ashwin Mathur，David Stap，Jay Gala，Wissam Siblini，Dominik Krzemi ´NSKI，NSKI，NSKI，NSKI，NSKI，GENTA INDRA WINATA等人。 MMTEB：大量的多语言文本嵌入基准。在
The Thirteenth International Conference on Learning Representations，2025。url https://openreview.net/forum?id=zl3pfz4VCV。 Tao Ge，Xin Chan，Xiaoyang Wang，Dian Yu，Haitao Mi和Dong Yu。用1,000,000,000个角色来扩展合成数据创建。
arXiv preprint arXiv:2406.20094，2024年。Jui-ting huang，Ashish Sharma，Shuying Sun，Li Xia，David Zhang，Philip Pronin，Janani Padman- Abhan，Giuseppe Ottaviano和Linjun Yang。在Facebook搜索中基于嵌入的检索。在
Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，第2553–2561页，2020年。 GPT-4O系统卡。 *arXiv preprint arXiv:2410.21276*，2024。VladimirKarpukhin，Barlas Oguz，Sewon Min，Patrick Sh Lewis，Ledell Wu，Sergey Edunov，Danqi Chen和Wen-Tau Yih。通态通道检索，以回答开放域的问题。在*EMNLP (1)*，第6769–6781页，2020年。ChankyuLee，Rajarshi Roy，Mengyao Xu，Jonathan Raiman，Mohammad Shoeybi，Bryan Catanzaro和Wei Ping。 NV Embed：作为通才嵌入模型训练LLM的改进技术。 *arXiv preprint arXiv:2405.17428*，2024年。ChankyuLee，Rajarshi Roy，Mengyao Xu，Jonathan Raiman，Mohammad Shoeybi，Bryan Catanzaro和Wei Ping。 NV Embed：作为通才嵌入模型训练LLM的改进技术。在
The Thirteenth International Conference on Learning Representations，2025a。 URL https://openreview.net/forum?id=lgsyLSsDRe。 Jinhyuk Lee，Feiyang Chen，Sahil Dua，Daniel Cer，Madhuri Shanbhogue，Iftekhar Naim，Gus-Tavo Hern ´Andez ´Andez ´Abrego，Zhe Li，Zhe Li，Kaifeng li，Kaifeng Chen，Henrique Schechter Vera等。双子座嵌入：双子座的可推广嵌入。
arXiv preprint arXiv:2503.07891，2025b。

Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035*, 2024.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023. URL https://arxiv.org/abs/2308.03281.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*, 2023.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1410/.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022. URL https://arxiv.org/abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.642/.

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. *arXiv preprint arXiv:2403.15246*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024. Association for Computing Machinery. URL https://doi.org/10.1145/3626772.3657878.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mingxin Li，Zhijie Nie，Yanzhao Zhang，Dingkun Long，Richong Zhang和Pengjun Xie。改善一般文本嵌入模型：通过模型合并解决任务冲突和数据不平衡。 *arXiv preprint arXiv:2410.15035*，2024年。迈向具有多阶段对比学习的一般文本嵌入，2023。URLhttps://arxiv.org/ abs/2308.03281。Xueguang MA，Xinyu Zhang，Ronak Pradeep和Jimmy Lin。用大语言模型重新播放零击文档。 *arXiv preprint arXiv:2305.02156*，2023年。NiklasMuennighoff，Nouamane Tazi，Loic Magne和Nils Reimers。 MTEB：大量嵌入基准的文本。在 *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*，第2014-2037页，2023年5月，克罗地亚杜布罗夫尼克。计算语言学协会。URL https://aclanthology.org/2023.eacl-main.148/。 Niklas Muennighoff，Hongjin Su，Liang Wang，Nan Yang，Furu Wei，Tao Yu，Amanpreet Singh和Douwe Kiela。生成代表性教学调整。在 *The Thirteenth International Con- ference on Learning Representations*，2025。url https://openreview.net/forum?id=BC4lIvfSzv中。亚伦·范·登·奥德（Aaron Van den Oord），Yazhe Li和Oriol Vinyals。表示对比度的谓语编码。 *arXiv preprint arXiv:1807.03748*，2018年。Ronakpradeep，Sahel SharifyMoghaddam和Jimmy Lin。 rankvicuna：零击列表文档使用开源大语模型。 *arXiv preprint arXiv:2309.15088*，2023。NilsReimers和Iryna Gurevych。句子 - 伯特：使用暹罗伯特网络的句子嵌入。在 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*，第3982–3992页，中国香港，2019年11月。计算语言学协会。 URL https://aclanthology.org/D19-1410/。 Hongjin Su，Weijia Shi，Jungo Kasai，Yizhong Wang，Yushi Hu，Mari Ostendorf，Wen-Tau Yih，Noah A Smith，Luke Zettlemoyer和Tao Yu。一个嵌入器，任何任务：指令 - 字段的文本嵌入。在 *Findings of the Association for Computational Linguistics: ACL 2023*，第1102–1121页，第2023页。文本嵌入通过弱监督的对比预训练，2022。URLhttps://arxiv.org/abs/2212.03533。 Liang Wang，Nan Yang，小黄，Linjun Yang，Rangan Majumder和Furu Wei。用大语言模型证明文本嵌入。在 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*，第11897–11916页，曼谷，泰国，2024年8月。计算语言学协会。 URL https://aclanthology.org/2024.acl-long.642/。 Orion Weller，Benjamin Chang，Sean Macavaney，Kyle LO，Arman Cohan，Benjamin van Durme，Dawn Lawrie和Luca Soldaini。关注：评估和教学信息检索模型以遵循说明。 *arXiv preprint arXiv:2403.15246*，2024年。ShitaoXiao，Zheng Liu，Peitian Zhang，Niklas Muennighoff，Defu Lian和Jian-Yun Nie。 C包：中国一般嵌入的包装资源。在 *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*，Sigir’24，第641–649页，纽约，纽约，美国，2024年。计算机协会。 URL https://doi.org/10.1145/ 3626772.3657878。 Yang，Anfeng Li，Baosong Yang，Beichen Zhang，Binyuan Hui，Bo Zheng，Bowen Yu，Chang Gao，Chengen Huang，Chenxu LV等。 QWEN3技术报告。 *arXiv preprint arXiv:2505.09388*，2025。

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. A two-stage adaptation of large language models for text ranking. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 11880–11891, 2024a.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412, Miami, Florida, US, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.103. URL https://aclanthology.org/2024.emnlp-industry.103/.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024.

Xiangyu Zhao, Maolin Wang, Xinjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. Embedding in recommender systems: A survey. *arXiv preprint arXiv:2310.18608*, 2023.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 38–47, 2024.

Longhui Zhang，Yanzhao Zhang，Dingkun Long，Pengjun Xie，Meishan Zhang和Min Zhang。大型语言模型进行文本排名的两阶段改编。在*Findings of the Association for Computational Linguistics ACL 2024*，第11880–11891页，第2024A页。

x 在Zhang，Yanzhao Zhang，Dingkun Long，Wen Xie，Ziqi Dai，Jialong Tang，Huan Lin，Huan Lin，Baosong Yang，Pengjun Xie，Fei Huang，Meishan Zhang，Wenjie Li和Min Zhang和Min Zhang。MGTE：多语言文本检索的广义长篇文本表示和重新脉动模型。在Franck Dernoncourt中，Daniel Preot ¸Iuc-Pietro和Anastasia Shimorina（编辑），*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*，第1393–1412页，迈阿密，佛罗里达，佛罗里达，美国，2024年11月，美国。计算语言学协会。 doi：10.18653/ v1 / 2024.emnlp-industry.103。 URL https://aclanthology.org/2024.emnlp-industry.103/。

韦恩Xin Zhao，Jing Liu，Ruiyang Ren和Ji-Rong Wen。基于预处理的语言模型的密集文本检索：调查。 *ACM Transactions on Information Systems*，42（4）：1–60，2024。

Xiangyu Zhao，Maolin Wang，Xinjian Zhao，Jiansheng Li，Shucheng Zhou，Dawei Yin，Qing Li，Jiliang Tang和Ruocheng Guo。嵌入推荐系统中：调查。 *arXiv preprint arXiv:2310.18608*，2023。

Shengyao Zhuang，Honglei Zhuang，Bevan Koopman和Guido Zuccon。具有大型语言模型的有效且高效的零摄像排名的盘点方法。在*Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*，第38–47、2024页。

A Appendix

A.1 Synthetic Data

We construct four types of synthetic data—retrieval, bitext mining, semantic textual similarity, and classification to enable the model to adapt to various similarity tasks during pre-training. To ensure both multilingual and cross-lingual diversity, the data is generated using Qwen3 32B. Below is an example of a synthetic retrieval text pair. The retrieval data is synthesized using a document-to-query approach. We collect a multilingual corpus from the pre-training corpus of the Qwen3 base model to serve as the document source. A two-stage generation pipeline is then applied, consisting of: (1) configuration and (2) query generation. In the configuration stage, we use large language models (LLMs) to determine the "Question Type", "Difficulty", and "Character" for the synthetic query. The candidate characters are retrieved from Persona Hub (Ge et al., 2024), selecting the top five most relevant to the given document. This step aims to enhance the diversity of the generated queries. The template used is as follows:

```
Given a **Passage** and **Character**, select the appropriate option from
 ↪  three fields: Character, Question_Type, Difficulty, and return the output
 ↪  in JSON format.
First, select the Character who are likely to be interested in the Passage
 ↪  from the candidates. Then select the Question_Type that the Character
 ↪  might ask about the Passage; Finally, choose the Difficulty of the
 ↪  possible question based on the Passage, the Character, and the
 ↪  Question_Type.
Character: Given by input **Character**

Question_Type:
- keywords: ...
- acquire_knowledge: ...
- summary: ...
- yes_or_no: ...
- background: ...

Difficulty:
- high_school: ...
- university: ...
- phd: ...

Here are some examples
<Example1> <Example2> <Example3>

Now, generate the **output** based on the **Passage** and **Character** from
 ↪  user, the **Passage** will be in {language} language and the **Character**
 ↪  will be in English.
Ensure to generate only the JSON output with content in English.

**Passage**:
{passage}
**Character**:
{character}
```

In the query generation stage, we use the configuration selected in the first stage to guide the generation of queries. Additionally, we explicitly specify the desired length and language of the generated query. The template used is as follows:

附录

A.1合成数据

我们构建了四种类型的合成数据：网状，bitext挖掘，语义文本相似性和分类，以使模型能够适应预训练期间的各种相似性任务。为了确保多语言和跨语性多样性，使用QWEN3 32B生成数据。以下是合成检索文本对的一个示例。使用文档到查询方法合成检索数据。我们从QWEN3基本模型的训练前语料库中收集多种语言语料库，以作为文档源。然后，应用了两阶段的管道，包括：（1）配置和（2）查询生成。在配置阶段，我们使用大型语言模型（LLMS）来确定合成查询的“问题类型”，“难度”和“字符”。候选人角色是从角色枢纽（Ge等，2024）中检索的，选择了与给定文档最相关的前五名。此步骤旨在增强生成的查询的多样性。所使用的模板如下：

```
Given a **Passage** and **Character**, select the appropriate option from
 ↪  three fields: Character, Question_Type, Difficulty, and return the output
 ↪  in JSON format.
First, select the Character who are likely to be interested in the Passage
 ↪  from the candidates. Then select the Question_Type that the Character
 ↪  might ask about the Passage; Finally, choose the Difficulty of the
 ↪  possible question based on the Passage, the Character, and the
 ↪  Question_Type.
Character: Given by input **Character**

Question_Type:
- keywords: ...
- acquire_knowledge: ...
- summary: ...
- yes_or_no: ...
- background: ...

Difficulty:
- high_school: ...
- university: ...
- phd: ...

Here are some examples
<Example1> <Example2> <Example3>

Now, generate the **output** based on the **Passage** and **Character** from
 ↪  user, the **Passage** will be in {language} language and the **Character**
 ↪  will be in English.
Ensure to generate only the JSON output with content in English.

**Passage**:
{passage}
**Character**:
{character}
```

在查询生成阶段，我们使用在第一阶段选择的配置来指导查询的生成。此外，我们明确指定生成的查询的所需长度和语言。所使用的模板如下：

```
Given a **Character**, **Passage**, and **Requirement**, generate a query from
 ↪  the **Character**'s perspective that satisfies the **Requirement** and can
 ↪  be used to retrieve the **Passage**. Please return the result in JSON
 ↪  format.

Here is an example:
<example>

Now, generate the **output** based on the **Character**, **Passage** and
 ↪  **Requirement** from user, the **Passage** will be in {corpus_language}
 ↪  language, the **Character** and **Requirement** will be in English.
Ensure to generate only the JSON output, with the key in English and the value
 ↪  in {queries_language} language.

**Character**
{character}
**Passage**
{passage}
**Requirment**
- Type: {type};
- Difficulty: {difficulty};
- Length: the length of the generated sentences should be {length} words;
- Languange: the language in which the results are generated should be
 ↪  {language} language;
```

Stage	Dataset	Size
Weakly Supervised Pre-Training	Synthetic Data	$\sim$ 150M
Supervised Fine Tuning	MS MARCO, NQ, HotpotQA, NLI, Dureader, T^2-Ranking, SimCLUE, MIRACL, MLDR, Mr.TyDi, Multi-CPR, CodeSearchNet .etc + High-quality Synthetic Data	Labeled Data: $\sim$ 7M Synthetic Data: $\sim$ 12M

Table 6: Statistics of training data utilized at each stage.

A.2 Detail Results

MTEB(eng, v2)	Param	Mean (Task)	Mean (Type)	Class-ification	Clus-tering	Pair Class.	Rerank	Retrieval	STS	Summ.
multilingual-e5-large-instruct	0.6B	65.53	61.21	75.54	49.89	86.24	48.74	53.47	84.72	29.89
NV-Embed-v2	7.8B	69.81	65.00	87.19	47.66	88.69	49.61	62.84	83.82	35.21
GritLM-7B	7.2B	67.07	63.22	81.25	50.82	87.29	49.59	54.95	83.03	35.65
gte-Qwen2-1.5B-instruct	1.5B	67.20	63.26	85.84	53.54	87.52	49.25	50.25	82.51	33.94
stella_en_1.5B_v5	1.5B	69.43	65.32	89.38	57.06	88.02	50.19	52.42	83.27	36.91
gte-Qwen2-7B-instruct	7.6B	70.72	65.77	88.52	58.97	85.9	50.47	58.09	82.69	35.74
gemini-embedding-exp-03-07	-	73.3	67.67	90.05	59.39	87.7	48.59	64.35	85.29	38.28
Qwen3-Embedding-0.6B	0.6B	70.70	64.88	85.76	54.05	84.37	48.18	61.83	86.57	33.43
Qwen3-Embedding-4B	4B	74.60	68.09	89.84	57.51	87.01	50.76	68.46	88.72	34.39
Qwen3-Embedding-8B	8B	75.22	68.70	90.43	58.57	87.52	51.56	69.44	88.58	34.83

Table 7: Results on MTEB(eng, v2) (Muennighoff et al., 2023). We compare models from the online leaderboard.

```
Given a **Character**, **Passage**, and **Requirement**, generate a query from
 ↪  the **Character**'s perspective that satisfies the **Requirement** and can
 ↪  be used to retrieve the **Passage**. Please return the result in JSON
 ↪  format.

Here is an example:
<example>

Now, generate the **output** based on the **Character**, **Passage** and
 ↪  **Requirement** from user, the **Passage** will be in {corpus_language}
 ↪  language, the **Character** and **Requirement** will be in English.
Ensure to generate only the JSON output, with the key in English and the value
 ↪  in {queries_language} language.

**Character**
{character}
**Passage**
{passage}
**Requirment**
- Type: {type};
- Difficulty: {difficulty};
- Length: the length of the generated sentences should be {length} words;
- Languange: the language in which the results are generated should be
 ↪  {language} language;
```

Stage	Dataset	Size
Weakly Supervised Pre-Training	Synthetic Data	$\sim$ 150M
Supervised Fine Tuning	MS MARCO, NQ, HotpotQA, NLI, Dureader, T^2-Ranking, SimCLUE, MIRACL, MLDR, Mr.TyDi, Multi-CPR, CodeSearchNet .etc + High-quality Synthetic Data	Labeled Data: $\sim$ 7M Synthetic Data: $\sim$ 12M

表6：在每个阶段使用的培训数据的统计数据。

A.2细节结果

MTEB(eng, v2)	Param	Mean (Task)	Mean (Type)	Class-ification	Clus-tering	Pair Class.	Rerank	Retrieval	STS	Summ.
multilingual-e5-large-instruct	0.6B	65.53	61.21	75.54	49.89	86.24	48.74	53.47	84.72	29.89
NV-Embed-v2	7.8B	69.81	65.00	87.19	47.66	88.69	49.61	62.84	83.82	35.21
GritLM-7B	7.2B	67.07	63.22	81.25	50.82	87.29	49.59	54.95	83.03	35.65
gte-Qwen2-1.5B-instruct	1.5B	67.20	63.26	85.84	53.54	87.52	49.25	50.25	82.51	33.94
stella_en_1.5B_v5	1.5B	69.43	65.32	89.38	57.06	88.02	50.19	52.42	83.27	36.91
gte-Qwen2-7B-instruct	7.6B	70.72	65.77	88.52	58.97	85.9	50.47	58.09	82.69	35.74
gemini-embedding-exp-03-07	-	73.3	67.67	90.05	59.39	87.7	48.59	64.35	85.29	38.28
Qwen3-Embedding-0.6B	0.6B	70.70	64.88	85.76	54.05	84.37	48.18	61.83	86.57	33.43
Qwen3-Embedding-4B	4B	74.60	68.09	89.84	57.51	87.01	50.76	68.46	88.72	34.39
Qwen3-Embedding-8B	8B	75.22	68.70	90.43	58.57	87.52	51.56	69.44	88.58	34.83

表7：MTEB（Eng，v2）的结果（Muennighoff等，2023）。我们比较在线排行榜中的模型。

MTEB(cmn, v1)	**Param**	**Mean (Task)**	**Mean (Type)**	Class-ification	Clus-tering	Pair Class.	Rerank	Retrieval	STS
multilingual-e5-large-instruct	0.6B	58.08	58.24	69.80	48.23	64.52	57.45	63.65	45.81
gte-Qwen2-7B-instruct	7.6B	71.62	72.19	75.77	66.06	81.16	69.24	75.70	65.20
gte-Qwen2-1.5B-instruct	1.5B	67.12	67.79	72.53	54.61	79.5	68.21	71.86	60.05
Qwen3-Embedding-0.6B	0.6B	66.33	67.44	71.40	68.74	76.42	62.58	71.03	54.52
Qwen3-Embedding-4B	4B	72.26	73.50	75.46	77.89	83.34	66.05	77.03	61.26
Qwen3-Embedding-8B	8B	73.84	75.00	76.97	80.08	84.23	66.99	78.21	63.53

Table 8: Results on C-MTEB (Xiao et al., 2024) (MTEB(cmn, v1).

MTEB(Code, v1)	Avg.	Apps	COIR-CodeSearch-Net	Code-Edit-Search	Code-Feedback-MT	Code-Feedback-ST	Code-SearchNet-CCR	Code-SearchNet	Code-Trans-Ocean-Contest	Code-Trans-Ocean-DL	CosQA	Stack-Overflow-QA	Synthetic-Text2SQL
$BGE_{multilingual}$	62.04	22.93	68.14	60.48	60.52	76.70	73.23	83.43	86.84	32.64	27.93	92.93	58.67
NV-Embed-v2	63.74	29.72	61.85	73.96	60.27	81.72	68.82	86.61	89.14	33.40	34.82	92.36	60.90
gte-Qwen2-7B-instruct	62.17	28.39	71.79	67.06	57.66	85.15	66.24	86.96	81.83	32.17	31.26	84.34	53.22
gte-Qwen2-1.5B-instruct	61.98	28.91	71.56	59.60	49.92	81.92	72.08	91.08	79.02	32.73	32.23	90.27	54.49
BGE-M3 (Dense)	58.22	14.77	58.07	59.83	47.86	69.27	53.55	61.98	86.22	29.37	27.36	80.71	49.65
Jina-v3	58.85	28.99	67.83	57.24	59.66	78.13	54.17	85.50	77.37	30.91	35.15	90.79	41.49
Qwen3-Embedding-0.6B	75.41	75.34	84.69	64.42	90.82	86.39	91.72	91.01	86.05	31.36	36.48	89.99	76.74
Qwen3-Embedding-4B	80.06	89.18	87.93	76.49	93.21	89.51	95.59	92.34	90.99	35.04	37.98	94.32	78.21
Qwen3-Embedding-8B	80.68	91.07	89.51	76.97	93.70	89.93	96.35	92.66	93.73	32.81	38.04	94.75	78.75
Qwen3-Reranker-0.6B	73.42	69.43	85.09	72.37	83.83	78.05	94.76	88.8	84.69	33.94	36.83	93.24	62.48
Qwen3-Reranker-4B	81.20	94.25	90.91	82.53	95.25	88.54	97.58	92.48	93.66	36.78	35.14	97.11	75.06
Qwen3-Reranker-8B	81.22	94.55	91.88	84.58	95.64	88.43	95.67	92.78	90.83	34.89	37.43	97.3	73.4

Table 9: Performance on `MTEB(Code, v1)` (Enevoldsen et al., 2025). We report nDCG@10 scores.

MTEB(cmn, v1)	**Param**	**Mean (Task)**	**Mean (Type)**	Class-ification	Clus-tering	Pair Class.	Rerank	Retrieval	STS
multilingual-e5-large-instruct	0.6B	58.08	58.24	69.80	48.23	64.52	57.45	63.65	45.81
gte-Qwen2-7B-instruct	7.6B	71.62	72.19	75.77	66.06	81.16	69.24	75.70	65.20
gte-Qwen2-1.5B-instruct	1.5B	67.12	67.79	72.53	54.61	79.5	68.21	71.86	60.05
Qwen3-Embedding-0.6B	0.6B	66.33	67.44	71.40	68.74	76.42	62.58	71.03	54.52
Qwen3-Embedding-4B	4B	72.26	73.50	75.46	77.89	83.34	66.05	77.03	61.26
Qwen3-Embedding-8B	8B	73.84	75.00	76.97	80.08	84.23	66.99	78.21	63.53

表8：C-MTEB的结果（Xiao等，2024）（MTEB（CMN，V1）。

MTEB(Code, v1)	Avg.	Apps	COIR-CodeSearch-Net	Code-Edit-Search	Code-Feedback-MT	Code-Feedback-ST	Code-SearchNet-CCR	Code-SearchNet	Code-Trans-Ocean-Contest	Code-Trans-Ocean-DL	CosQA	Stack-Overflow-QA	Synthetic-Text2SQL
$BGE_{multilingual}$	62.04	22.93	68.14	60.48	60.52	76.70	73.23	83.43	86.84	32.64	27.93	92.93	58.67
NV-Embed-v2	63.74	29.72	61.85	73.96	60.27	81.72	68.82	86.61	89.14	33.40	34.82	92.36	60.90
gte-Qwen2-7B-instruct	62.17	28.39	71.79	67.06	57.66	85.15	66.24	86.96	81.83	32.17	31.26	84.34	53.22
gte-Qwen2-1.5B-instruct	61.98	28.91	71.56	59.60	49.92	81.92	72.08	91.08	79.02	32.73	32.23	90.27	54.49
BGE-M3 (Dense)	58.22	14.77	58.07	59.83	47.86	69.27	53.55	61.98	86.22	29.37	27.36	80.71	49.65
Jina-v3	58.85	28.99	67.83	57.24	59.66	78.13	54.17	85.50	77.37	30.91	35.15	90.79	41.49
Qwen3-Embedding-0.6B	75.41	75.34	84.69	64.42	90.82	86.39	91.72	91.01	86.05	31.36	36.48	89.99	76.74
Qwen3-Embedding-4B	80.06	89.18	87.93	76.49	93.21	89.51	95.59	92.34	90.99	35.04	37.98	94.32	78.21
Qwen3-Embedding-8B	80.68	91.07	89.51	76.97	93.70	89.93	96.35	92.66	93.73	32.81	38.04	94.75	78.75
Qwen3-Reranker-0.6B	73.42	69.43	85.09	72.37	83.83	78.05	94.76	88.8	84.69	33.94	36.83	93.24	62.48
Qwen3-Reranker-4B	81.20	94.25	90.91	82.53	95.25	88.54	97.58	92.48	93.66	36.78	35.14	97.11	75.06
Qwen3-Reranker-8B	81.22	94.55	91.88	84.58	95.64	88.43	95.67	92.78	90.83	34.89	37.43	97.3	73.4

表9：`MTEB(Code, v1)`（Enevoldsen等人，2025）上的性能。我们报告NDCG@10分。