本文由 AINLP 公众号整理翻译，更多 LLM 资源请扫码关注!

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排序

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 实验室 阿里巴巴集团

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

摘要

在本工作中，我们推出了 Qwen3 Embedding 系列，这是在其前身 GTE-Qwen 系列基础上的重大提升，具有更强的文本嵌入和重排序能力，建立在 Qwen3 基础模型之上。利用 Qwen3 大型语言模型在多语言文本理解和生成方面的强大能力，我们创新的多阶段训练流程结合了大规模无监督预训练与在高质量数据集上的有监督微调。有效的模型融合策略进一步确保了 Qwen3 Embedding 系列的稳健性和适应性。在训练过程中，Qwen3 大型语言模型不仅作为骨干模型，还在跨多个领域和语言合成高质量、丰富且多样的训练数据中发挥关键作用，从而增强了训练流程。Qwen3 Embedding 系列提供多种模型规模（0.6B、4B、8B），适用于嵌入和重排序任务，满足不同部署场景，用户可以根据效率或效果进行优化。实证评估显示，Qwen3 Embedding 系列在各种基准测试中实现了最先进的结果。值得注意的是，它在多语言评估基准 MTEB 的文本嵌入任务中表现出色，以及在包括代码检索、跨语言检索和多语言检索在内的多种检索任务中表现优异。为了促进可复现性并推动社区驱动的研究与开发，Qwen3 Embedding 模型已在 Apache 2.0 许可证下公开提供。

1 引言

文本嵌入和重排序是众多自然语言处理和信息检索应用中的基础组成部分，包括网页搜索、问答系统、推荐系统及其他（Karpukhin 等，2020；Huang 等，2020；Zhao 等，2023；2024）。高质量的嵌入使模型能够捕捉文本之间的语义关系，而有效的重排序机制则确保最相关的结果被优先展示。近年来，随着大型语言模型（如 Qwen3（Yang 等，2025）、GPT-4o（Hurst 等，2024））的进步，出现了一些新兴的应用范式，如检索增强生成（RAG）和智能代理系统，这些都对文本嵌入和重排序提出了新的需求和挑战，无论是在模型训练范式还是在应用场景方面。尽管取得了显著的进展，但训练在可扩展性、上下文理解和与特定下游任务的对齐方面表现良好的嵌入和重排序模型仍然具有挑战性。

大型语言模型（LLMs）的出现极大地推动了文本嵌入和重排序模型的发展。在引入LLMs之前，主要的方法

* 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

涉及使用仅编码器的预训练语言模型，如 BERT，作为训练的基础模型（Reimers & Gurevych, 2019）。LLMs 内在的丰富世界知识、文本理解和推理能力，推动了在这些架构上训练的模型的进一步提升。此外，在将 LLMs 集成到训练数据合成和高质量数据筛选等流程中的研究也取得了显著进展（Wang 等，2024；Lee 等，2024；2025b）。LLMs 的基本特性也激发了新训练范式的引入。例如，在嵌入模型训练过程中，结合指令类型、领域和语言等方面的差异化任务，提升了下游任务的性能（Su 等，2023）。类似地，在重排序模型训练中，通过基于用户提示的零样本方法和结合监督微调的方法（Ma 等，2023；Pradeep 等，2023；Zhang 等，2024a；Zhuang 等，2024）实现了进步。

在本工作中，我们引入了Qwen3嵌入系列模型，这些模型是在Qwen3基础模型之上构建的。Qwen3基础模型同时发布了基础版和指令版，我们利用这些模型在多语言文本理解和生成方面的强大能力，充分发挥其在训练嵌入和重排序模型中的潜力。为了训练嵌入模型，我们实现了一个多阶段的训练流程，包括大规模的无监督预训练，然后在高质量数据集上进行有监督的微调。我们还采用模型合并的方法，结合不同的模型检查点，以增强模型的鲁棒性和泛化能力。Qwen3指令模型能够高效合成大量高质量、多语言、多任务的文本相关性数据集。这些合成数据在初始的无监督训练阶段被利用，而一部分高质量的小规模数据则被选用于第二阶段的有监督训练。对于重排序模型，我们采用类似的两阶段训练方案，包括高质量的有监督微调和模型合并阶段。基于不同规模的Qwen3骨干模型（包括0.6B、4B和8B），我们最终训练了三种文本嵌入模型和三种文本重排序模型。为了方便其在下游任务中的应用，Qwen3嵌入系列支持多项实用功能，例如嵌入模型的灵活维度表示和嵌入及重排序模型的可定制指令。

我们在涵盖多个任务和领域的全面基准测试中评估了Qwen3 Embedding系列。实验结果表明，我们的嵌入和重排序模型实现了最先进的性能，在多个检索任务中与领先的专有模型竞争。例如，旗舰模型Qwen3-8B-Embedding在多语种基准测试（Enevoldsen等，2025）中获得了70.58的分数，在MTEB代码基准测试（Enevoldsen等，2025）中获得了80.68的分数，超越了之前的最先进的专有嵌入模型Gemini-Embedding（Lee等，2025b）。此外，我们的重排序模型在一系列检索任务中也表现出具有竞争力的结果。Qwen3-Reranker-0.6B模型在许多检索任务中超过了之前表现最好的模型，而更大的Qwen3-Reranker-8B模型表现更优，在多个任务中比0.6B模型的排名结果提升了3.0分。此外，我们还进行了具有建设性的消融研究，以阐明促成Qwen3 Embedding系列优越性能的关键因素，为其有效性提供了见解。

在接下来的章节中，我们将描述模型架构的设计，详细介绍训练过程，展示Qwen3嵌入系列的嵌入和重排序模型的实验结果，并通过总结关键发现和提出未来研究的潜在方向来结束本技术报告。

2 模型架构

嵌入和重排序模型的核心思想是以任务感知的方式评估相关性。给定一个查询 q 和一个文档 d，嵌入和重排序模型根据由指令 I 定义的相似性标准来评估它们的相关性。为了使模型能够进行任务感知的相关性估计，训练数据通常被组织为 $\{I_i、q_i、d_i^+、d_{i,1}^-、\cdots、d_{i,n}^-\}$，其中 d_i^+ 代表一个

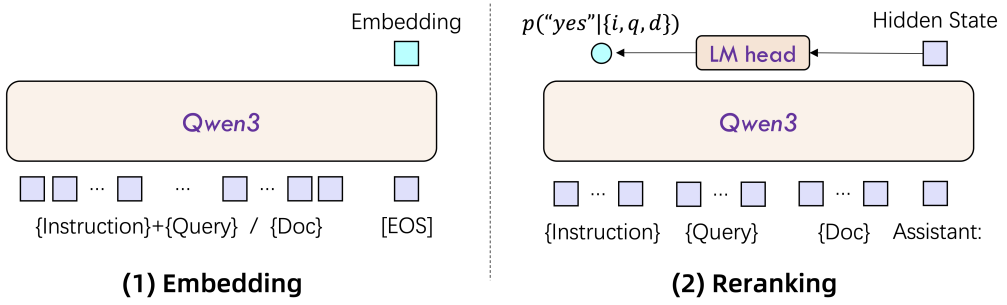

Figure 1: Model architecture of Qwen3-Embedding (left) and Qwen3-Reranker (right).

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
```

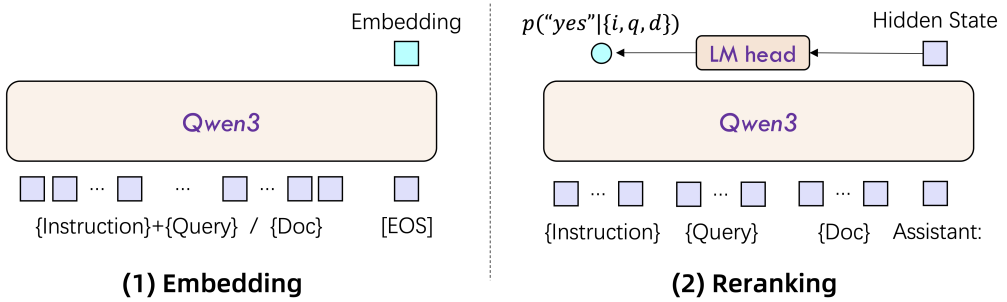

图 1：Qwen3-嵌入模型（左）和 Qwen3-重排序器（右）的模型架构。

与查询 q_i 相关的正面（相关）文档，以及 $d_{i,j}^-$ 为负面（无关）文档。在多样化文本对上训练模型可以拓宽其在一系列下游任务中的适用性，包括检索、语义文本相似度、分类和聚类。

架构 Qwen3 嵌入和重排序模型建立在 Qwen3 基础模型的稠密版本之上，提供三种尺寸：0.6B、4B 和 8B 参数。我们使用 Qwen3 基础模型进行初始化，以利用其在文本建模和指令执行方面的能力。每个模型配置的模型层数、隐藏层大小和上下文长度详见表1。

嵌入模型 对于文本嵌入，我们使用带有因果注意力的大型语言模型（LLMs），在输入序列的末尾添加一个 `[EOS]` 标记。最终的嵌入是由对应于这个 `[EOS]` 标记的最后一层隐藏状态得出的。

为了确保嵌入在下游任务中遵循指令，我们将指令和查询连接成一个输入上下文，同时在使用大型语言模型（LLMs）处理之前保持文档不变。查询的输入格式如下：

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

重新排序模型 为了更准确地评估文本相似度，我们在单一上下文中采用LLMs进行点对点的重新排序。与嵌入模型类似，为了实现指令跟随能力，我们在输入上下文中加入指令。我们使用LLM聊天模板，并将相似度评估任务框架为二分类问题。输入到LLMs的内容遵循如下模板：

```
<|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.

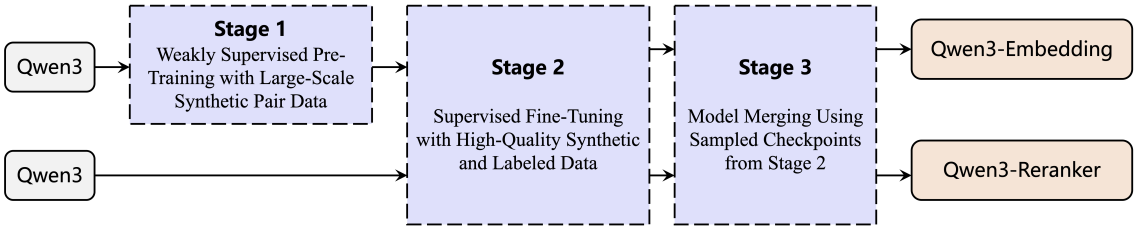

Figure 2: Training pipeline of Qwen3 Embedding and Reranking models.

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:

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

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嵌入模型的架构。“MRL支持”表示嵌入模型是否支持自定义最终嵌入的维度。“指令感知”指出嵌入或重排序模型是否支持根据不同任务自定义输入指令。

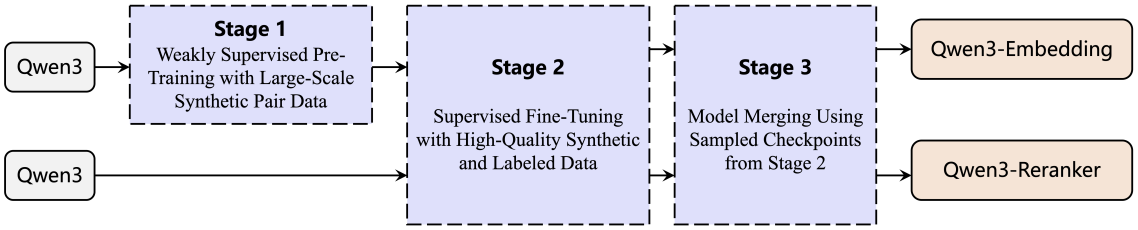

图2：Qwen3嵌入和重排序模型的训练流程。

为了根据给定输入计算相关性得分，我们评估下一个标记为“是”或“否”的可能性。其数学表达式如下：

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

3 模型训练

在本节中，我们描述所采用的多阶段训练流程，并介绍该训练方案的关键要素，包括训练目标、训练数据合成以及高质量训练数据的筛选。

3.1 训练目标

在介绍我们的训练流程之前，首先概述在训练过程中用于嵌入和重排序模型的优化损失函数。对于嵌入模型，我们采用基于InfoNCE框架（Oord等，2018）的改进对比损失。给定一批N训练实例，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 $s(\cdot,\cdot)$ 是相似度函数（我们使用余弦相似度），τ 是温度参数，Z_i 是归一化因子，用于将正样本对的相似度分数与各种负样本对的相似度分数进行聚合：

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

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 positive and negative documents d_j. 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.

其中这些项表示与以下内容的相似度：(1) 正样本文档 d_i^+，(2) K 硬负样本 $d_{i,k}^-$，(3) 批次中的其他查询 q_j，(4) 批次中的其他正负样本文档 d_j。掩码因子 m_{ij} 旨在减轻假负样本的影响，其定义为：

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

其中 s_{ij} 是 q_i、d_j 或 q_i、q_j 的对应得分。

对于重排序模型，我们优化定义为：监督微调（SFT）损失。

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

其中 $p(\cdot|*)$ 表示由大型语言模型（LLM）赋予的概率。标签 l 对于正面文档为“是”，对于负面文档为“否”。此损失函数鼓励模型为正确的标签赋予更高的概率，从而提升排序性能。

3.2 多阶段训练

多阶段训练方法是训练文本嵌入模型的常用做法（Li 等，2023；Wang 等，2022；Chen 等，2024）。该策略通常首先在包含噪声的大规模半监督数据上进行初步训练，然后使用较小的高质量监督数据集进行微调。这一两步过程提升了嵌入模型的性能和泛化能力。大规模的弱监督训练数据对模型的泛化能力有重要贡献，而在后续阶段用高质量数据进行微调则进一步改善模型性能。嵌入模型的两个训练阶段都采用了在方程 1 中定义的优化目标，而重排序模型的训练则以在方程 2 中定义的损失函数作为优化目标。

在现有多阶段训练框架的基础上，Qwen3 Embedding 系列引入了以下关键创新：

- 大规模合成数据驱动的弱监督训练：与以往的工作（例如 GTE、E5、BGE 模型）不同，后者主要从开源社区如问答论坛或学术论文中收集弱监督训练数据，我们提出利用基础模型的文本理解和生成能力，直接合成配对数据。这种方法允许在合成提示中任意定义所需配对数据的各种维度，如任务、语言、长度和难度。与从开放领域源收集数据相比，基础模型驱动的数据合成具有更强的可控性，能够精确管理生成数据的质量和多样性，特别是在资源有限的场景和语言中。

- 在有监督微调中的高质量合成数据利用：由于Qwen3基础模型的卓越性能，合成数据的质量显著提高。因此，在有监督训练的第二阶段，选择性地引入这些高质量的合成数据，进一步提升了整体模型的性能和泛化能力。

- 模型合并：受到之前工作的启发（Li 等，2024），在完成有监督的微调后，我们采用了一种基于球面线性插值（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。此步骤旨在提升模型在各种数据分布下的鲁棒性和泛化性能。

需要注意的是，重排序模型的训练过程不包括第一阶段的弱监督训练阶段。

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

为了创建一个稳健的合成数据集，用于训练模型在各种相似性任务上，我们生成了涵盖检索、比对挖掘、分类和语义文本相似性（STS）等类别的多样化文本对。这些合成数据对的质量通过使用Qwen3-32B模型作为数据合成的基础模型得以保证。我们设计了一套多样化的提示策略，以提高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进行数据合成。在数据合成过程中，为每个文档分配特定角色，以模拟潜在用户对该文档的查询。这种引入用户视角的方法增强了合成查询的多样性和真实性。具体而言，我们利用检索模型从角色库中为每个文档识别出前五个角色候选，并将这些文档及其角色候选呈现给提示。这引导模型输出最适合的角色配置以生成查询。此外，提示还结合了查询类型（例如，关键词、事实、摘要、判断）、查询长度、难度和语言等多个维度。这种多维度的方法确保了合成数据的质量和多样性。

最后，我们共创建了大约1.5亿对多任务弱监督训练数据。我们的实验显示，使用这些合成数据训练的嵌入模型在下游评估中表现尤为出色，特别是在MTEB多语言基准测试中超越了许多之前的有监督模型。这激励我们对合成数据进行筛选，以识别高质量的对，并在第二阶段的有监督训练中纳入。我们采用简单的余弦相似度计算来选择数据对，从随机采样的数据中保留余弦相似度大于0.7的对。最终，约有1200万对高质量的有监督训练数据被选中用于进一步训练。

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：在MTEB多语种（Enevoldsen等，2025）上的表现。对于对比模型，分数取自2025年6月4日的MTEB在线排行榜。

4 评估

我们在多个基准测试中进行全面且公平的评估，以评估Qwen3 Embedding模型的能力。

4.1 设置

对于文本嵌入模型，我们使用大规模多语言文本嵌入基准（MMTEB）（Enevoldsen 等，2025）进行评估。MMTEB 是对 MTEB（Muennighoff 等，2023）的一个由社区推动的扩展，涵盖了超过 500 个经过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：在MTEB英语、MTEB中文、MTEB代码上的表现。[α]取自（Enevoldsen等，2025年）。[γ]取自（Lee等，2025b）。对于其他对比模型，分数取自2025年6月4日的MTEB在线排行榜。

涵盖超过250种语言。除了经典的文本任务，如各种检索、分类和语义文本相似度外，MMTEB还包括一系列具有挑战性和新颖性的任务，例如指令遵循、长文档检索和代码检索，代表迄今为止最大规模的多语言嵌入模型评估任务集合。我们的MMTEB评估涵盖216个单独的评估任务，包括131个MTEB（多语言）任务（Enevoldsen等，2025）、41个MTEB（英语，v2）任务（Muennighoff等，2023）、32个CMTEB任务（Xiao等，2024）以及12个MTEB（代码）代码检索任务（Enevoldsen等，2025）。

此外，我们选择了一系列文本检索任务来评估我们模型的文本重排序能力。我们探索了三种类型的检索任务：(1) 基本相关性检索，分为英语、中文和多语言，在 MTEB (Muennighoff et al., 2023)、CMTEB (Xiao et al., 2024)、MMTEB (Enevoldsen et al., 2025) 和 MLDR (Chen et al., 2024) 上进行评估；(2) 代码检索，在 MTEB-Code (Enevoldsen et al., 2025) 上进行评估，该数据集仅包含与代码相关的检索数据；以及 (3) 复杂指令检索，在 FollowIR (Weller et al., 2024) 上进行评估。

对比方法 我们将我们的模型与最突出的开源文本嵌入模型和商业API服务进行比较。开源模型包括GTE（Li等，2023；Zhang等，2024b）、E5（Wang等，2022）和BGE（Xiao等，2024）系列，以及NV-Embed-v2（Lee等，2025a）、GritLM-7B Muennighoff等（2025）。评估的商业API包括OpenAI的text-embedding-3-large、Google的Gemini-embedding，以及Cohere-embed-multilingual-v3.0。对于重排序，我们与jina[1]的重排序器、mGTE（Zhang等，2024b）和BGE-m3（Chen等，2024）进行了比较。

4.2 主要结果

嵌入 在表2中，我们展示了在MMTEB（Enevoldsen等，2025）上的评估结果，该数据集全面涵盖了多种语言的广泛嵌入任务。我们的Qwen3-Embedding-4B/8B模型实现了最佳性能，而我们最小的模型Qwen3-Embedding-0.6B，尽管只有0.6B参数，但仅次于表现最好的基线方法（Gemini-Embedding）。在表3中，我们展示了在MTEB（英语，v2）（Muennighoff等，2023）、CMTEB（Xiao等，2024）和MTEB（Code）（Enevoldsen等，2025）上的评估结果。评分反映出与MMTEB类似的趋势，我们的Qwen3-Embedding-4B/8B模型

[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：重排序模型的评估结果。我们使用了MTEB(eng, v2)、MTEB(cmn, v1)和MMTEB的检索子集，分别为MTEB-R、CMTEB-R和MMTEB-R。其余均为检索任务。所有得分均为我们基于第一行检索前100名结果的运行结果。

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：Qwen3-Embedding-0.6B模型在不同训练设置下，在MMTEB、MTEB（英语，v2）、CMTEB和MTEB（代码，v1）上的性能（平均任务）

始终优于其他模型。值得注意的是，Qwen3-Embedding-0.6B模型仅次于Gemini-Embedding，并且在与gte-Qwen2-7B-instruct的竞争中具有一定的优势。

重新排序 在表4中，我们展示了在各种重新排序任务（§4.1）上的评估结果。我们使用Qwen3-Embedding-0.6B模型检索前100个候选项，然后应用不同的重新排序模型进行进一步优化。这种方法确保了对重新排序模型的公平评估。我们的结果表明，所有三个Qwen3-Reranker模型都优于嵌入模型，并且超越了所有基线重新排序方法，其中Qwen3-Reranker-8B在大多数任务中都取得了最高的性能。

4.3 分析

为了进一步分析和探索Qwen3嵌入模型训练框架的关键要素，我们从以下几个维度进行分析：

大规模弱监督预训练的有效性 我们首先分析大规模弱监督训练阶段对嵌入模型的效果。如表5所示，仅在合成数据上训练的Qwen3-Embedding-0.6B模型（没有后续的训练阶段，如第一行所示）在性能上表现合理且较强，与最终的Qwen3-Embedding-0.6B模型（如最后一行所示）相比。如果我们进一步去除弱监督训练阶段（即不进行合成数据训练，如第二行所示），最终性能明显下降。这表明，大规模弱监督训练阶段对于实现优越的性能至关重要。

模型合并的效果 接下来，我们比较模型合并阶段带来的性能差异。如表5所示，不使用模型合并技术（第三行，使用数据采样平衡各种任务）训练的模型，其性能明显不如最终的Qwen3-Embedding-0.6B模型（采用模型合并，如最后一行所示）。这表明模型合并阶段对于开发强大模型也至关重要。

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

在本技术报告中，我们介绍了 Qwen3-Embedding 系列，这是一套基于 Qwen3 基础模型的全面文本嵌入和重排序模型。这些模型旨在在广泛的文本嵌入和重排序任务中表现出色，包括多语言检索、代码检索和复杂指令执行。Qwen3-Embedding 模型建立在一个强大的多阶段训练流程之上，结合了在合成数据上的大规模弱监督预训练、在高质量数据集上的监督微调以及模型融合。Qwen3 大型语言模型在合成跨多语言和多任务的多样化训练数据方面发挥了关键作用，从而增强了模型的能力。我们的全面评估显示，Qwen3-Embedding 模型在多个基准测试中实现了最先进的性能，包括 MTEB 、CMTEB、MMTEB 以及若干检索基准。我们很高兴开源 Qwen3-Embedding 和 Qwen3-Reranker 模型（0.6B、4B 和 8B），使社区能够使用和基于它们进行开发。

参考文献

陈建宇，肖石涛，张佩天，罗坤，连德福，刘正。在
Findings of the Association for Computational Linguistics: ACL 2024，第 2318–2335 页，泰国曼谷，2024 年 8 月。计算语言学协会。网址 https:// aclanthology.org/2024.findings-acl.137/。Kenneth Enevoldsen，Isaac Chung，Imene Kerboua，Márton Kardos，Ashwin Mathur，David Stap，Jay Gala，Wissam Siblini，Dominik Krzemi ński，Genta Indra Winata 等。MMTEB：大规模多语言文本嵌入基准。在 *The Thirteenth International Conference on Learning Representations*，2025 年。网址 https://openreview.net/forum?id=zl3pfz4VCV。Ge Tao，Xin Chan，Wang Xiaoyang，Yu Dian，Mi Haitao，余东。利用 10 亿个角色扩展合成数据创建。*arXiv preprint arXiv:2406.20094*，2024 年。黄瑞廷，Sharma Ashish，Sun Shuying，Xia Li，Zhang David，Pronin Philip，Padman- abhan Janani，Ottaviano Giuseppe，杨林军。Facebook 搜索中的基于嵌入的检索。在
Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，第 2553–2561 页，2020 年。Aaron Hurst，Lerer Adam，Goucher Adam P，Perelman Adam，Ramesh Aditya，Clark Aidan，Ostrow AJ，Welihinda Akila，Hayes Alan，Radford Alec 等。GPT-4o 系统卡片。*arXiv preprint arXiv:2410.21276*，2024 年。Vladimir Karpukhin，Barlas Oguz，Min Sewon，Lewis Patrick SH，Wu Ledell，Edunov Sergey，Chen Danqi，Yih Wen-tau。面向开放域问答的密集段落检索。在 *EMNLP (1)*，第 6769–6781 页，2020 年。李灿奎，Rajarshi Roy，Xu Mengyao，Raiman Jonathan，Shoeybi Mohammad，Catanzaro Bryan，Ping Wei。Nv-embed：用于训练通用嵌入模型的改进技术。*arXiv preprint arXiv:2405.17428*，2024 年。李灿奎，Rajarshi Roy，Xu Mengyao，Raiman Jonathan，Shoeybi Mohammad，Catanzaro Bryan，Ping Wei。NV-embed：用于训练通用嵌入模型的改进技术。在 *The Thirteenth International Conference on Learning Representations*，2025a 年。网址 https://openreview.net/forum?id=lgsyLSsDRe。李振煜，陈飞扬，Dua Sahil，Cer Daniel，Shanbhogue Madhuri，Naim Iftekhar，Gustavo Hernández Ábrego，Li Zhe，Chen Kaifeng，Vera Henrique Schechter，等。Gemini 嵌入：来自 Gemini 的可泛化嵌入。*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.

明新李、志杰聂、燕昭张、丁坤龙、荣聪张、彭俊谢。改进通用文本嵌入模型：通过模型融合应对任务冲突和数据不平衡。*arXiv preprint arXiv:2410.15035*，2024年。李泽涵、张鑫、燕昭张、丁坤龙、彭俊谢、张梅珊。基于多阶段对比学习的通用文本嵌入方法，2023年。网址 https://arxiv.org/abs/2308.03281。马学光、张新宇、Ronak Pradeep 和 Jimmy Lin。利用大型语言模型进行零-shot列表式文档重排序。*arXiv preprint arXiv:2305.02156*，2023年。尼克拉斯·穆恩尼霍夫、Nouamane Tazi、Loic Magne 和 Nils Reimers。MTEB：大规模文本嵌入基准。在 *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*，第2014-2037页，克罗地亚杜布罗夫尼克，2023年5月。计算语言学协会。网址 https://aclanthology.org/2023.eacl-main.148/。尼克拉斯·穆恩尼霍夫、苏宏金、王亮、杨楠、魏福如、于涛、Amanpreet Singh 和 Douwe Kiela。生成式表示指令调优。在 *The Thirteenth International Con- ference on Learning Representations*，2025年。网址 https://openreview.net/forum?id=BC4lIvfSzv。Aaron van den Oord、Li Yazhe 和 Oriol Vinyals。对比预测编码的表示学习。*arXiv preprint arXiv:1807.03748*，2018年。Ronak Pradeep、Sahel Sharifymoghaddam 和 Jimmy Lin。RankVicuña：利用开源大型语言模型进行零-shot列表式文档重排序。*arXiv preprint arXiv:2309.15088*，2023年。尼尔斯·雷默斯和伊琳娜·古雷维奇。Sentence-BERT：使用孪生BERT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)*，2019年11月，香港。计算语言学协会。网址 https://aclanthology.org/D19-1410/。苏宏金、史伟佳、笠井淳吾、王一中、胡玉石、奥斯特福德、易文涛、史诺亚·A·史密斯、Luke Zettlemoyer 和于涛。一种嵌入器，适用于任何任务：指令微调的文本嵌入。在 *Findings of the Association for Computational Linguistics: ACL 2023*，第1102-1121页，2023年。王亮、杨楠、黄晓龙、焦宾学、杨林俊、江大鑫、Rangan Majumder 和魏福如。通过弱监督对比预训练的文本嵌入，2022年。网址 https://arxiv.org/abs/2212.03533。王亮、杨楠、黄晓龙、杨林俊、Rangan Majumder 和魏福如。利用大型语言模型提升文本嵌入。 在 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*，第11897-11916页，泰国曼谷，2024年8月。计算语言学协会。网址 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。FollowIR：评估和指导信息检索模型遵循指令。*arXiv preprint arXiv:2403.15246*，2024年。肖志涛、刘正、张培天、Niklas Muennighoff、连德富 和聂建云。C-pack：面向通用中文嵌入的资源包。在 *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*，SIGIR ’24，第641-649页，纽约，美国，2024年。计算机协会。网址 https://doi.org/10.1145/3626772.3657878。安阳、李安峰、杨宝松、张贝辰、惠彬源、郑博、于博、高昌、黄承根、吕辰旭 等。Qwen3技术报告。*arXiv preprint arXiv:2505.09388*，2025年。

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.

长辉张、燕昭张、丁坤龙、鹏军谢、梅山张、敏张。大规模语言模型的两阶段适应用于文本排序。发表于 *Findings of the Association for Computational Linguistics ACL 2024*，第 11880–11891 页，2024 a。

X在张艳昭、张丁坤、龙文、谢子奇、戴佳龙、唐焕、林宝松、杨鹏军、谢飞、黄梅山、张文杰、李敏等人。mGTE：用于多语言文本检索的广义长上下文文本表示和重排序模型。收录于Franck Dernoncourt、Daniel Preoţiuc-Pietro 和 Anastasia Shimorina（编），*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*，第1393–1412页，佛罗里达州迈阿密，2024年11月。计算语言学协会。doi: 10.18653/v1/2024.emnlp-industry.103。网址 https://aclanthology.org/2024.emnlp-industry.103/。

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, 和 Ji-Rong Wen。基于预训练语言模型的密集文本检索：一项综述。*ACM Transactions on Information Systems*，42(4):1–60，2024。

赵向宇，王茂林，赵新建，李建胜，周树成，尹大伟，李青，唐纪良，郭若澄。推荐系统中的嵌入：一项综述。*arXiv preprint arXiv:2310.18608*，2023。

庄胜耀，庄宏磊，库普曼 Bevan 和 Zuccon Guido. 一种用于大规模语言模型的高效零样本排序的集合方法。在 *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*，第 38–47 页，2024 年。

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:

A 附录

A.1 合成数据

我们构建了四种类型的合成数据——检索、双语文本挖掘、语义文本相似性和分类，以使模型在预训练期间能够适应各种相似性任务。为了确保多语言和跨语言的多样性，数据采用Qwen3 32B生成。以下是一个合成检索文本对的示例。检索数据采用文档到查询的方法进行合成。我们从Qwen3基础模型的预训练语料库中收集多语言语料，作为文档来源。然后应用两阶段生成流程，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“角色”。候选角色从Persona Hub（Ge等人，2024）中检索，选择与给定文档最相关的前五个角色。此步骤旨在增强生成查询的多样性。所用模板如下：

```
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，我们明确指定生成查询的期望长度和语言。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：每个阶段使用的训练数据统计。

A.2 详细结果

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）上的结果（Muennighoff 等，2023）。我们比较了在线排行榜上的模型。

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（Xiao 等，2024）（MTEB(cmn, v1)）的结果

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：在 `MTEB(Code, v1)` (Enevoldsen 等人，2025) 上的性能。我们报告 nDCG@10 分数。