

Qwen3 Embedding：提升文本嵌入和重排序能力

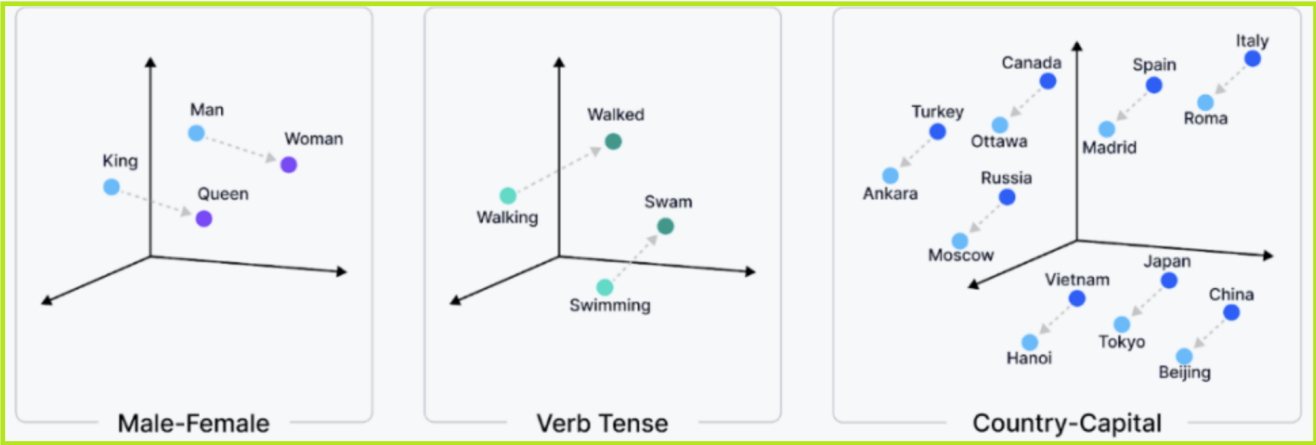
1. 论文：Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models

<https://arxiv.org/abs/2506.05176>
2. 博客地址： <https://qwenlm.github.io/blog/qwen3-embedding/>

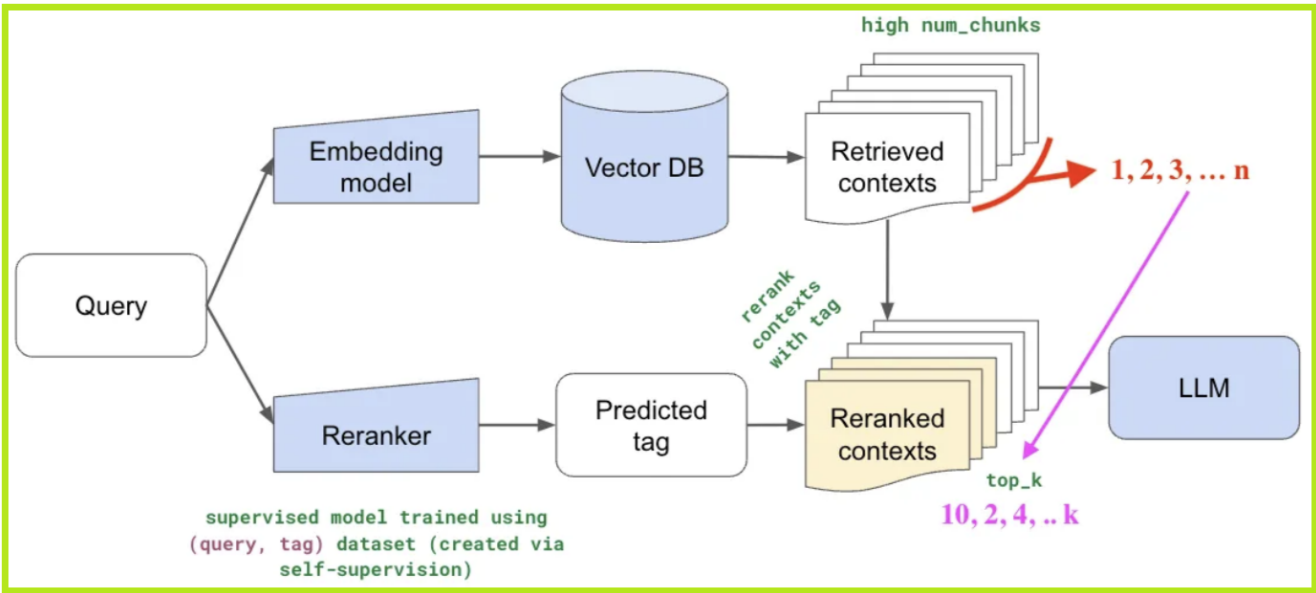
1. 研究背景与动机

在自然语言处理（NLP）和信息检索（IR）领域，文本嵌入（Text Embedding）和文本重排序（Reranking）是基石技术。

- 文本嵌入：将文本（词语、句子、文档）转换为能够捕捉其语义信息的数值向量。可以想象成在“语义空间”中给每段文本一个坐标，相似的文本在空间中距离更近。

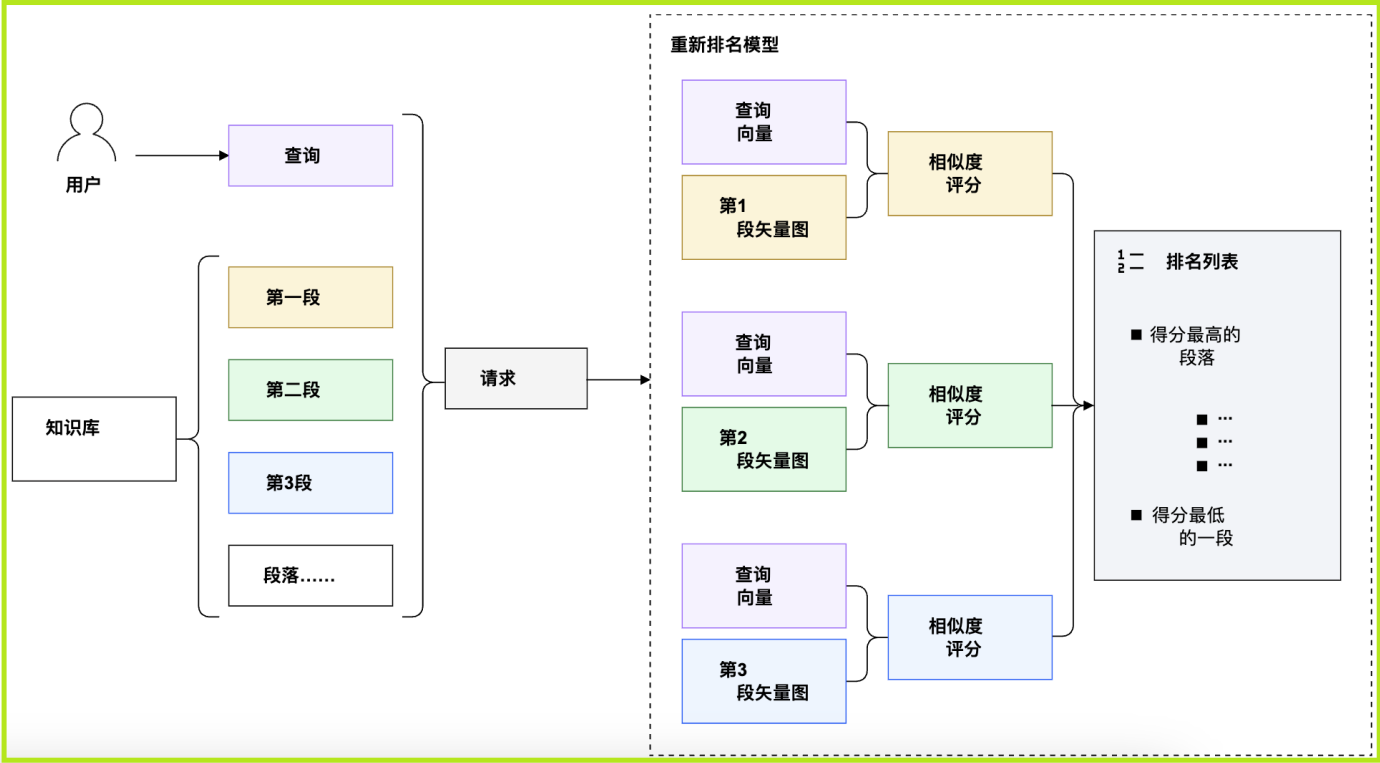


- 文本重排序：当我们为一个查询检索到一组候选文档后，重排序模型会更仔细地审查这些候选文档，将最相关的结果排在最前面。



这些技术对于以下应用至关重要：

- 网页搜索
- 问答系统（QA）
- 推荐系统
- 以及日益重要的新兴范式，如检索增强生成（RAG）和智能体（Agent）系统，它们都严重依赖于查找相关信息以输入给大语言模型（LLM）。



2. 论文贡献

Qwen3 Embedding系列模型带来了以下几个关键创新点：

1. **基于基础模型构建**：依托强大的**Qwen3基础模型**（参数规模包括0.6B、4B、8B），充分利用其在多语言文本理解和生成方面的鲁棒能力，用于训练嵌入和重排序模型。



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.

2. **创新的多阶段训练流程**：

- 结合了大规模无监督预训练（基于合成数据）和高质量数据集上的监督微调。
- 引入了基于不同检查点的模型合并（model merging）技术（使用slerp），以增强模型的鲁棒性和泛化能力。

```
# 模型权重合并（开源项目）
https://github.com/arcee-ai/mergekit.git
```

- 3.  LLM驱动的数据合成：一个显著特点是利用Qwen3 - 32B 本身来合成了大规模、高质量、多样化的多领域、多语言训练数据。这极大地增强了训练流程的效果。
- 4.  达到SOTA性能：在多个基准测试中取得了当前最佳（State-of-the-Art, SOTA）的结果，尤其在：
 - 多语言文本嵌入基准 MTEB (Massive Text Embedding Benchmark) 上表现优异。
 - 在代码检索、跨语言检索和多语言检索等任务中表现出色。

	Qwen3- Embedding-8B	Qwen3- Embedding-4B	Qwen3- Embedding-0.6B	Gemini Embedding	Cohere-embed- multilingual- v3.0	text- embedding-3- large	multilingual- e5-large- instruct	gte-Qwen2- 7B-instruct
MMTEB <small>Mean-Task</small>	70.58	69.45	64.33	68.37	61.12	58.93	63.22	62.51
MTEB (en v2) <small>Mean-Task</small>	75.22	74.60	70.70	73.30	66.01	66.43	65.53	70.72
MTEB-Code	80.68	80.06	75.41	74.66	51.94	58.95	65.00	56.41

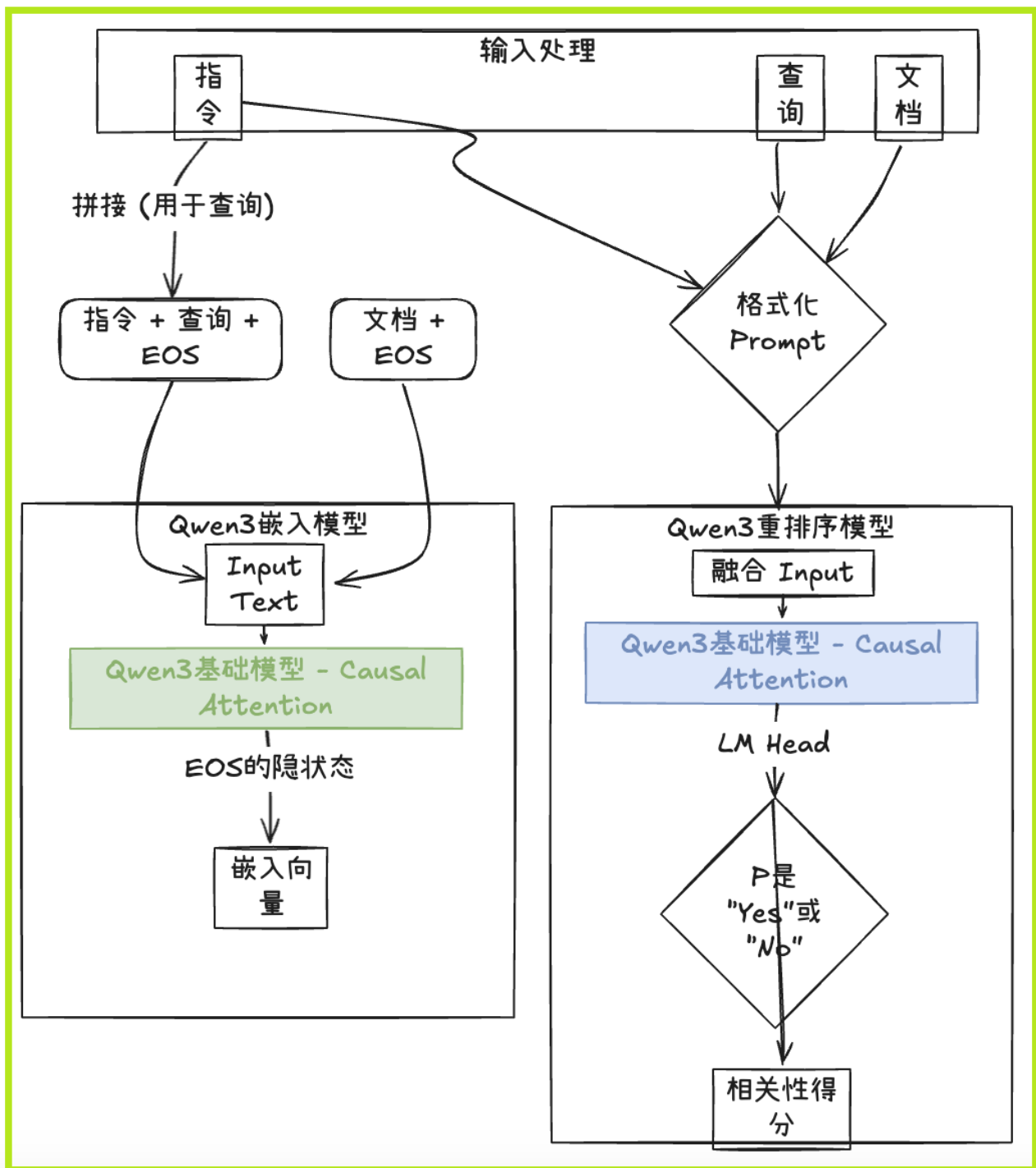
3. 核心算法原理

3.1 模型架构

Qwen3 Embedding系列模型利用了Qwen3基础模型，这些是Decoder-only的LLM。

1. 嵌入模型 (Embedding Models):

- 使用具有因果注意力（Causal Attention）的LLM。
- 在输入序列末尾附加一个 [EOS] (End Of Sequence) 标记，最终的嵌入向量来源于最后一层对应于此 [EOS] 标记的隐状态。
- 指令遵循 (Instruction Following): 为了使嵌入具备任务感知能力，指令会与查询拼接。
 - 查询输入格式: {Instruction} {Query}<|endoftext|>
 - 文档输入格式: {Document}<|endoftext|> (文档独立处理)



2. 重排序模型 (Reranking Models):

- 利用LLM进行逐点重排序 (**point-wise reranking**) (一次评估一个查询-文档对)。
- 任务被构建为一个二分类问题: 预测“yes” (相关) 或“no” (不相关)。
- 输入遵循LLM的聊天模板, 包含指令、查询和文档。

```
<|im_start|>system
根据Query和提供的Instruct判断Document是否满足要求。注意，答案只能是"yes"或"no"。<|im_end|>
<|im_start|>user
<Instruct>: {Instruction}
<Query>: {Query}
<Document>: {Document}<|im_end|>
<|im_start|>assistant
<think>\n\n</think>\n\n
```

- 相关性得分基于下一个词元是“yes”与“no”的似然比。

3. 模型配置 (Table 1):

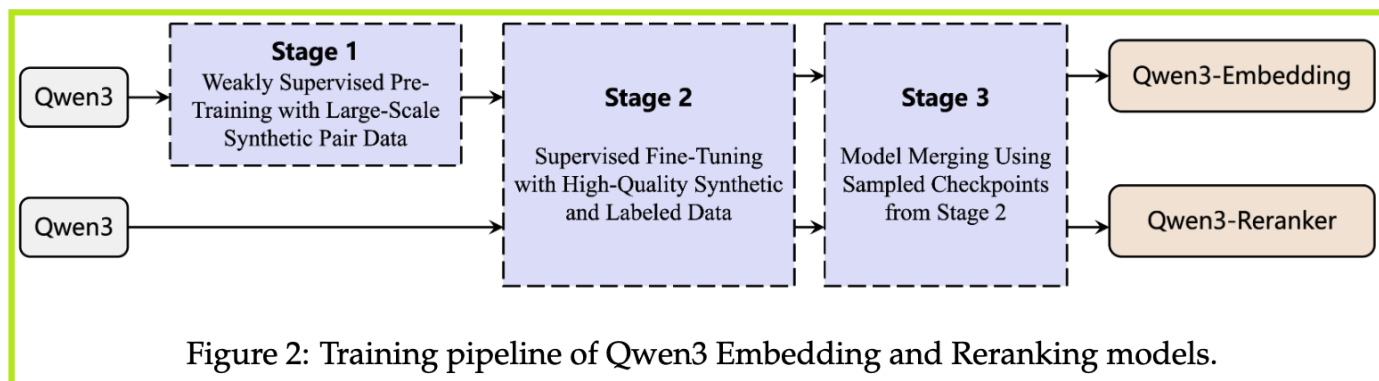
- 模型大小: 0.6B, 4B, 8B。
- 序列长度: 32K。
- 嵌入维度: 1024 (0.6B), 2560 (4B), 4096 (8B)。
- 所有模型都具备“指令感知 (Instruction Aware)”能力。嵌入模型支持“MRL Support”（自定义维度）。
 - 指令感知 (Instruction Aware):** 无论是Embedding还是Reranking模型，都支持用户根据不同任务定制输入指令，使模型能够更好地适应特定场景的需求。
 - 多表示层支持 (MRL Support):** Embedding模型支持自定义最终输出的嵌入维度，这为不同应用场景下的效率和存储需求提供了灵活性。

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.

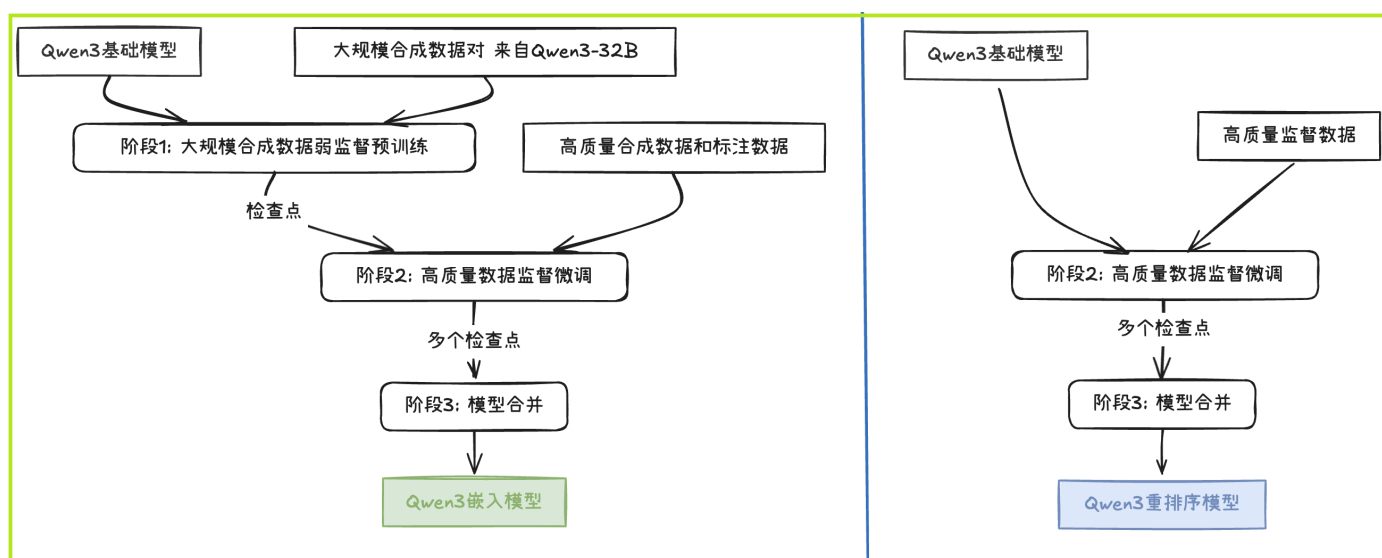
3.2 训练流程

采用多阶段训练方法 (如论文Fig.2所示)。



关键阶段与创新点：

- **阶段1: 大规模合成数据驱动的弱监督预训练 (仅嵌入模型)**
 - 与以往依赖问答论坛等数据的工作不同，Qwen3利用其强大的Qwen3-32B模型直接合成了约1.5亿个文本对。
- **阶段2: 监督微调 (嵌入和重排序模型)**
 - 使用规模较小但质量更高的数据集。
 - 关键在于，除了精心筛选的人工标注数据集（约700万对），还选择性地引入了高质量的合成数据（从阶段1筛选得到，通过余弦相似度大于0.7的标准筛选出约1200万对高质量数据）。
- **阶段3: 模型合并 (嵌入和重排序模型)**
 - 在监督微调完成后，采用基于球面线性插值（slerp）的模型融合技术，合并微调过程中保存的多个模型检查点，以提升模型的鲁棒性和泛化能力。
 - **Slerp简单理解**: 想象球面上有两个点（代表两个检查点的模型权重）。Slerp会找到这两个点之间沿球面最短的路径。你可以选择这条路径上的某个中间点作为融合后的模型。



3.3 训练目标

1. 嵌入模型 (基于InfoNCE的对比损失):

InfoNCE 是一种对比学习中的损失函数，用于最大化正样本之间的相似度，同时最小化与负样本之间的相似度。

目标是在嵌入空间中拉近正样本对（查询，相关文档）的距离，同时推开负样本对（查询，不相关文档）的距离。

给定一个包含 N 个训练实例的批次，损失函数定义为：

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

最小化 L_i 等价于最大化正样本对相对于所有负样本的对数概率。

其中：

- $s(\cdot, \cdot)$ 是余弦相似度函数。
- q_i 是查询 i 的嵌入。
- d_i^+ 是查询 i 对应的正样本（相关）文档的嵌入。
- τ 是温度参数。
- Z_i 是归一化因子，计算方式为对以下各项的得分求和：

$$Z_i = \underbrace{e^{s(q_i, d_i^+)/\tau}}_{\text{第1项: 正样本}} + \underbrace{\sum_{k=1}^K m_{ik} e^{s(q_i, d_{ik}^-)/\tau}}_{\text{第2项: 硬负样本}} + \underbrace{\sum_{j \neq i} m_{ij} e^{s(q_i, q_j)/\tau}}_{\text{第3项: 同批次内其他查询作为负样本}} + \underbrace{\sum_{j \neq i} m_{ij} e^{s(d_i^+, d_j)/\tau}}_{\text{第4项: 同批次内其他文档作为 } d_i^+ \text{ 的负样本}}$$

1. 正样本对: $e^{s(q_i, d_i^+)/\tau}$

2. K 个硬负样本 d_{ik}^- (对于查询 q_i): $\sum_k m_{ik} e^{s(q_i, d_{ik}^-)/\tau}$

3. 批次内其他查询 q_j (作为负样本): $\sum_{j \neq i} m_{ij} e^{s(q_i, q_j)/\tau}$

4. 批次内其他正样本 d_j^+ (作为 d_i^+ 的负样本): $\sum_{j \neq i} m_{ij} e^{s(d_i^+, d_j)/\tau}$

- m_{ij} 是一个掩码因子，旨在减轻伪负样本（false negatives）的影响（例如，如果批次内的 d_j 实际上与 q_i 非常相似）。

2. 重排序模型 (监督微调损失):

这是一个标准的分类损失：

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

监督微调的目标是最大化模型预测出正确标签的概率，这通常通过最小化负对数似然损失 (Negative Log-Likelihood Loss) 来实现。

其中：

- l 是真实标签 ("yes" 或 "no")。
- $P(q, d)$ 表示输入给LLM的内容 (查询 q , 文档 d , 指令 I)。
- $p(l|P(q, d))$ 是LLM赋给正确标签的概率。

4. 实验设计与结果

4.1 评估设置

- **主要基准 (嵌入模型): MMTEB (Massive Multilingual Text Embedding Benchmark)**
 - 覆盖超过250种语言，包含超过500个评估任务。
 - 评估的子集包括:
 - MTEB 多语言 (131个任务)
 - MTEB 英语 v2 (41个任务)
 - CMTEB (中文, 32个任务)
 - MTEB 代码 (12个代码检索任务)
- **重排序任务:**
 - 基础相关性检索: MTEB, CMTEB, MMTEB的检索子集, MLDR。
 - 代码检索: MTEB-Code。
 - 复杂指令检索: FollowIR。
- **对比方法 (SOTA模型):**
 - 开源模型: GTE, E5, BGE系列, NV-Embed-v2, GritLM-7B。
 - 商业API: OpenAI text-embedding-3-large, Gemini-Embedding, Cohere-embed-multilingual-v3.0。
 - 重排序器: jina-reranker, mGTE (reranker), BGE-m3 (reranker)。

4.2 主要结果

嵌入模型:

- **MTEB 多语言 (Table 2):**
 - **Qwen3-Embedding-8B:** 获得 **70.58**分，超过了 Gemini Embedding (68.37) 和之前的SOTA。
 - **Qwen3-Embedding-4B:** 得分 **69.45**，同样优于 Gemini。
 - **Qwen3-Embedding-0.6B:** 得分 **64.33**，在同等规模下极具竞争力。

Model	Size	Mean (Task)	Mean (Type)	Bitext Mining	Classification	Clustering	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.

• MTEB 英语, 中文, 代码 (Table 3):

- MTEB 英语 (v2): Qwen3-Embedding-8B (75.22) 优于 Gemini (73.30)。
- CMTEB (中文): Qwen3-Embedding-8B (73.83) 表现非常强劲。
- MTEB 代码:
 - Qwen3-Embedding-8B: **80.68** (显著领先于Gemini的74.66)。
 - Qwen3-Embedding-4B: **80.06**。
 - Qwen3-Embedding-0.6B: **75.41** (已超过Gemini)。
 这突显了模型卓越的代码嵌入能力。

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 English, MTEB Chinese, MTEB Code. ^αTaken from (Enevoldsen et al., 2025). ^γTaken from (Lee et al., 2025b). For other compared models, the scores are retrieved from MTEB online leaderboard on June 4th, 2025.

重排序模型 (Table 4):

(使用 Qwen3-Embedding-0.6B 进行初步 top-100 检索)

- 所有三个Qwen3-Reranker模型 (0.6B, 4B, 8B) 均显著提升了基准嵌入模型的性能，它们超越了所有基线重排序方法。
- **Qwen3-Reranker-8B** 在大多数任务上通常取得最高性能。
 - 例如：在MMTEB-R上, Qwen3-Embedding-0.6B (64.64) -> Qwen3-Reranker-0.6B (66.36) -> Qwen3-Reranker-4B (72.74) -> Qwen3-Reranker-8B (72.94)。
 - 在FollowIR (复杂指令)上, Qwen3-Reranker-4B 的提升幅度非常大 (14.84)。

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 MMTEB-R. The rest are all retrieval tasks. All scores are our runs based on the retrieval top-100 results from the first row.

4.3 消融实验

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.

- 大规模弱监督预训练 (阶段1) 的有效性:
 - 仅使用合成数据训练 (仅阶段1): 58.49 (MMTEB得分)
 - 不使用合成数据预训练 (仅阶段2和3): 61.21 (MMTEB得分)
 - 完整模型 (所有阶段): 64.33 (MMTEB得分)
 - **⚠️ 结论:** 移除弱监督预训练阶段会导致性能明显下降。仅在合成数据上训练效果尚可，但并非SOTA。三者结合才是关键。
- 模型合并 (阶段3) 的有效性:
 - 不进行模型合并 (阶段1和2): 62.56 (MMTEB得分)
 - 完整模型 (所有阶段，包含合并): 64.33 (MMTEB得分)
 - **⚠️ 结论:** 模型合并带来了显著的性能提升，表明其对于提升模型的鲁棒性和泛化能力非常重要。

5. 总结

Qwen3 Embedding这篇论文做出了重要贡献：

- 展示了如何有效地将强大的基础LLM (Qwen3) 改编用于专门的文本嵌入和重排序任务。
- 引入了一个创新的训练流程，特色在于LLM驱动的合成数据生成和模型合并。
- 提供了SOTA的开源模型，在多语言和代码相关场景中表现尤为突出。