

Qwen3 Embedding + Reranking Demo

这个 Demo 演示了如何结合使用 Qwen3 Embedding 和 Reranker 模型来实现高效的文档检索和重排序系统。

一、应用场景

- 知识库问答系统: 快速从大型知识库中检索相关文档
- 搜索引擎: 提高搜索结果的相关性和准确性
- 推荐系统: 为用户推荐最相关的内容
- 文档检索系统: 企业内部文档管理和检索

二、Demo 架构

两阶段检索架构

用户查询 → Embedding 粗排 → Reranker 精排 → 最终结果

1. 第一阶段 - Embedding 粗排:

- 使用 Embedding 模型将查询和文档转换为向量
- 通过余弦相似度计算快速筛选候选文档
- 优势: 速度快, 可以处理大规模文档库
- 缺点: 精度相对较低

2. 第二阶段 - Reranker 精排:

- 使用 Reranker 模型对候选文档进行精确评分
- 考虑查询和文档之间的深层语义关系
- 优势: 精度高, 相关性判断更准确
- 缺点: 计算开销较大, 适合处理少量候选文档

三、快速开始

环境要求

```
# 1. 安装 uv
# 使用官方安装脚本
curl -LsSf https://astral.sh/uv/install.sh | sh
# 或者通过 pip 安装
pip install uv

# 2. 在根目录下执行命令, 创建虚拟环境 .qwen
uv venv .qwen

# 3. 激活虚拟环境
source .qwen/bin/activate

# 4. 安装依赖包
uv pip install vllm==0.9.0 transformers==4.52.4 sentence-transformers==4.1.0 modelscope -i
https://pypi.tuna.tsinghua.edu.cn/simple

# 5. 安装 flash-attn
```

```
# pip install flash-attn --no-build-isolation

# 模型下载
# 候选: 0.6B , 4B , 8B
modelscope download --model Qwen/Qwen3-Embedding-0.6B --local_dir /root/autodl-tmp/models/Qwen3-Embedding-0.6B

# 候选: 0.6B, 4B, 8B
modelscope download --model Qwen/Qwen3-Reranker-0.6B --local_dir /root/autodl-tmp/models/Qwen3-Reranker-0.6B
```

运行 Demo

```
cd examples
python embedding_reranking_demo.py
```

四、Demo 功能

1. 自动测试模式

Demo 会自动运行 2个预定义的测试查询:

- 中国的首都是哪里?
- 什么是重力?

2. 交互式查询模式

测试完成后, 可以手动输入查询进行实时搜索。

3. 性能监控

- 显示 Embedding 和 Reranking 阶段的耗时
- 展示各阶段的分数变化
- 对比重排序前后的结果差异

AI进阶学习

 B站课堂: <https://space.bilibili.com/3546748815411393/pugv>

 官方网站: <https://www.tgltommy.com/> (国内访问需科学上网)

如果你正在关注大模型 Agent、强化学习后训练、RLHF、DPO、GRPO、RLVR 等前沿方向, 欢迎关注我最新上线的精品课程:

[《大模型 Agent 强化学习实战: 从后训练、奖励设计到工业级对齐系统》](#)

这门课程围绕当前大模型 Agent 强化学习的核心技术路线展开, 不只是介绍零散算法和论文概念, 而是希望帮助学习者建立一套完整的技术地图: 从基础原理、论文脉络、关键算法, 到开源项目源码解析与工业级系统实践, 逐步理解 Agent RL 为什么重要、如何演进, 以及在真实应用中如何落地。

课程配套专属飞书知识库《Agent RL 学习宝典》, 包含学习路线、论文资料、项目说明、源码阅读辅助资料和后续持续更新内容, 方便长期查阅和跟进前沿技术。

Agent RL 变化很快, 但越是变化快, 越需要一套系统的学习框架。如果你希望系统进入 Agent RL 方向, 而不是停留在碎片化了解阶段, 这门课会是一个很好的起点。

课程官网: <https://www.tgltommy.com/p/agent-rl> (国内访问需科学上网)

B站课堂: <https://www.bilibili.com/cheese/play/ss842375604>

更多课程详情, 扫描以下二维码咨询课程助理:

新课首发

大模型 Agent 强化学习实战:

从后训练、奖励设计到工业级对齐系统

唐国梁Tommy

TGL AI Tommy

官方微信课程助理

B站/公众号/YouTube: 唐国梁Tommy

**关注微信公众号
获取更多学习资源**



官方网站: TGLTommy.com(需科学上网)