科大讯飞 iFLYTEK | 开发者大赛 AI DEVELOPER COMPETITION

2025科大讯飞AI开发者大赛

面向公开和垂直领域的混域检索挑战赛

决赛答辩

答辩人：尹雅博

目录

CONTENTS

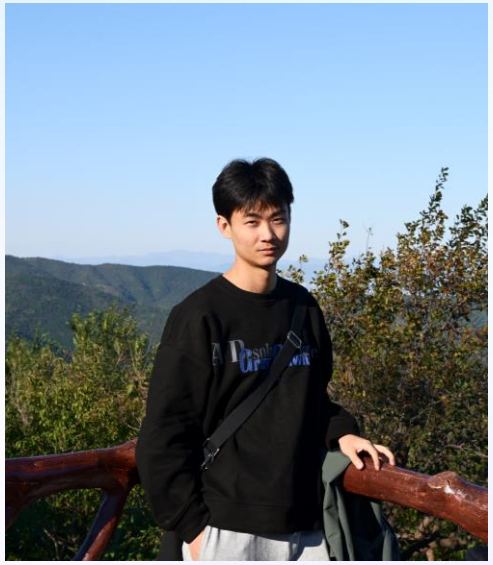
01 团队介绍

InsiTek

团队成员

尹雅博
毕业于重庆理工大学

功能实现,思路整理



奚阳
就读于北京交通大学

社区方案调研, 文档整理

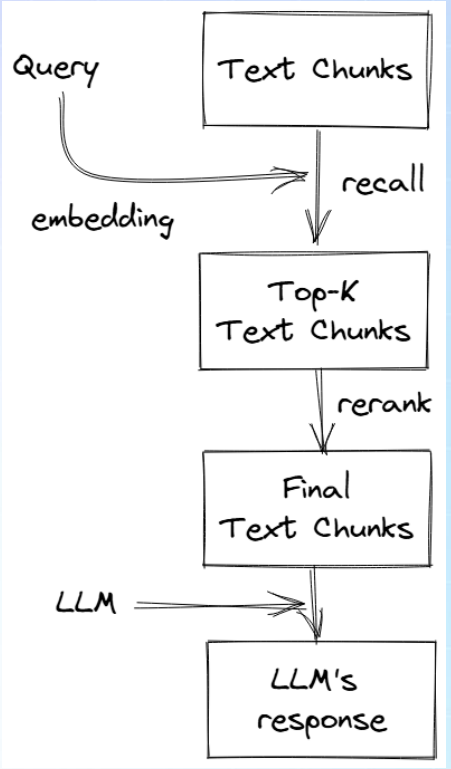
02 算法方案解析

问题背景:

当前大语言模型在垂直领域应用中,仍面临“幻觉”与“知识滞后”两大关键瓶颈,影响结果可靠性. 检索增强生成(RAG)通过结合信息检索来从外部知识库中检索相关信息,已成为大模型落地的关键路径:

- 增强模型知识覆盖能力
- 缓解知识更新滞后问题
- 提升问答、摘要等任务的准确性与可信度

尽管在RAG的设计上涌现出一系列优秀的方案, 如CRAG[1]、GraphRAG[2]、Agentic RAG[3]等,但在垂域问答场景中,检索模块的精准性仍是制约整体性能的关键因素,直接影响最终回答质量.



Simple RAG pipeline

任务定义:

本赛题基于讯飞互联网搜索结果的数据以及私有化的文档知识数据来构建一个Reranker模型,以支持对用户查询与给定给定文档（Document）或网页内容(Web Content) 的相关性判断.
输入：Query + Document 或 Query + Web Content
输出：是否相关的二分类判别

[1] Yan S Q, Gu J C, Zhu Y, et al. Corrective retrieval augmented generation[J]. 2024.
[2] Edge D, Trinh H, Cheng N, et al. From local to global: A graph rag approach to query-focused summarization[J]. arXiv preprint arXiv:2404.16130, 2024.
[3] Singh A, Ehtesham A, Kumar S, et al. Agentic retrieval-augmented generation: A survey on agentic rag[J]. arXiv preprint arXiv:2501.09136, 2025.

02 算法方案解析

核心问题:

在解决实际场景下的混域检索时，面临两大核心挑战：

通用重排序模型泛化能力不足.

通用 Reranker（如 BGE-M3-Reranker、Qwen3-Reranker等）在通用语料上进行训练，难以捕捉垂直领域的专业术语、逻辑结构和语义匹配模式，导致在实际应用中性能不佳.

垂直领域模型训练数据构建困难，标注数据噪声高.

垂直领域标注成本高，常依赖弱监督、众包、基于大模型的数据合成等数据构建方法，导致训练集中存在大量错误标签（如不相关样本被标为相关）.

为了解决上述问题,我们提出了一种**两阶段的协同优化框架，结合数据清洗、模型微调与推理校准实现垂直领域下的混域检索.**

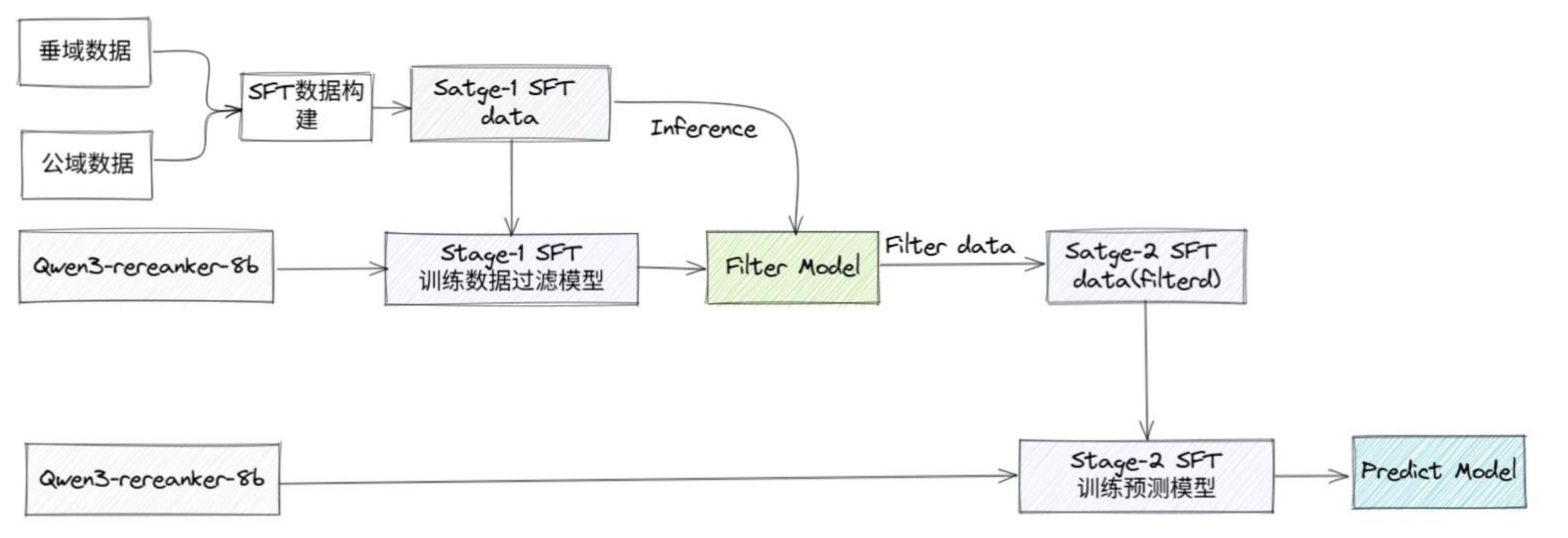
'query': '中国铁路跨入高铁时代的标志是什么？',

'doc_content': '知行天下高速铁路问答.txt\n海外项目在建设过程中存在与国外技术标准兼容互通的问题一是技术标准兼容性的问题；二是设备兼容性的问题；三是设计理念与技术方法联系并存的问题。\n《高速铁路设计--基础设施》标准和《高速铁路设计--供电》标准 2022年8月，国际铁路联盟（UIC)发布实施。 由中国国家铁路集团有限公司组织专家主持，法国、德国、日本、西班牙、意大利等十余个国家的20余名专家参与，历时4年编制而成。\n怎么看待中国高铁技术标准 一方面，中国需要在一定程度上向国际标准靠拢，来防止一些市场准入壁垒；另一方面，中国主要是依据国内高铁建设得来的经验建立中国高铁技术标准，“走出去”就难免会遇到各种问题。\n高铁文化 产业报国，促中国速度勇攀高峰。高铁产业技术、人才和资本高度密集，中国高铁能够在很短时间内从“并跑”到“领跑”世界先进水平，关键是高铁人从院士到一线工人，都有把国家利益、集体利益放在至高地位的报国情怀。 锲而不舍，科技攻关永不言弃。高铁成就背后，有无数人甘于充当“无名英雄”。',

'label': 'yes'

数据集中的错误标注样本case

02 算法方案解析

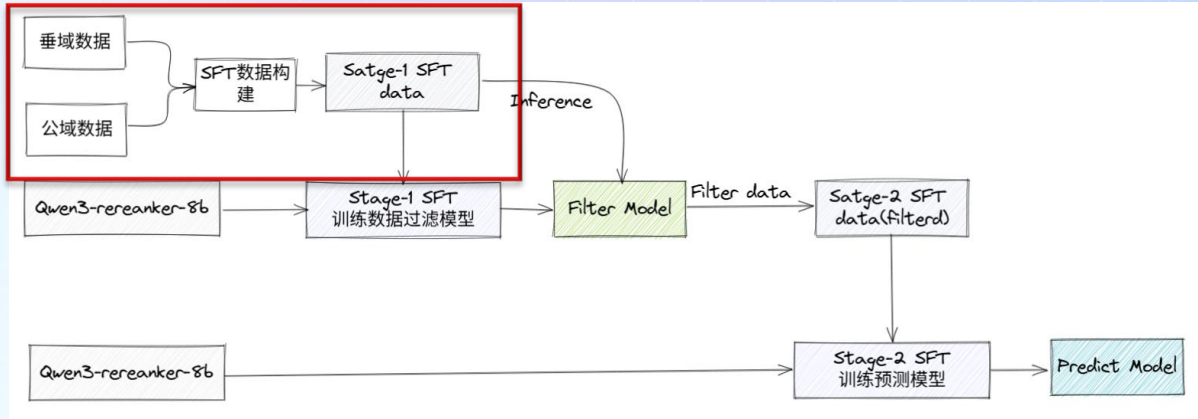


整体方案的Pipeline

02 算法方案解析

数据构建:

对于公域和垂域的数据,我们使用统一的方式来构建Alpaca-Style的SFT数据.



垂域 instruct_format

```
{
    '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".',
    'instruction': '<Instruct>: Given a web search query, retrieve relevant passages that answer the query\n\n<Query>: ##query##\n\n<Document>: ##document_content##',
    'input': '',
    'output': 'yes/no'
}
```

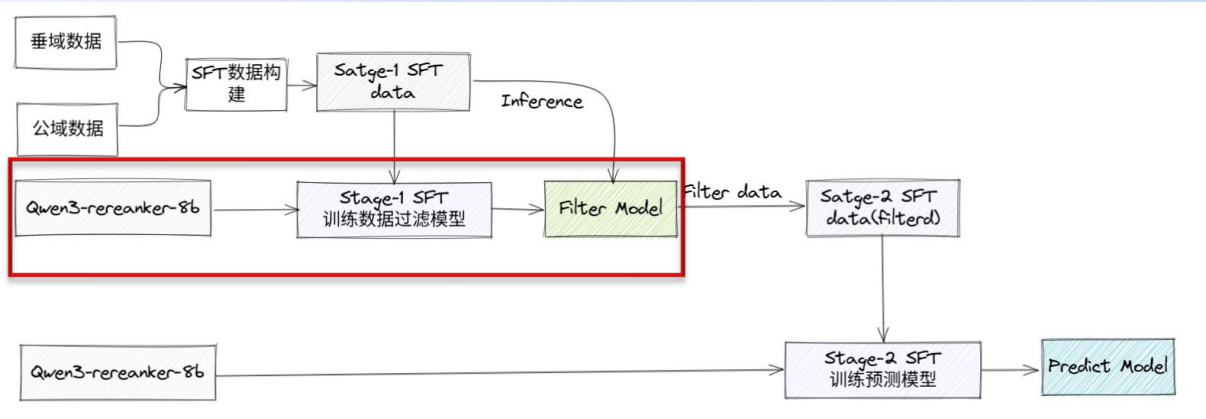
公域 instruct_format

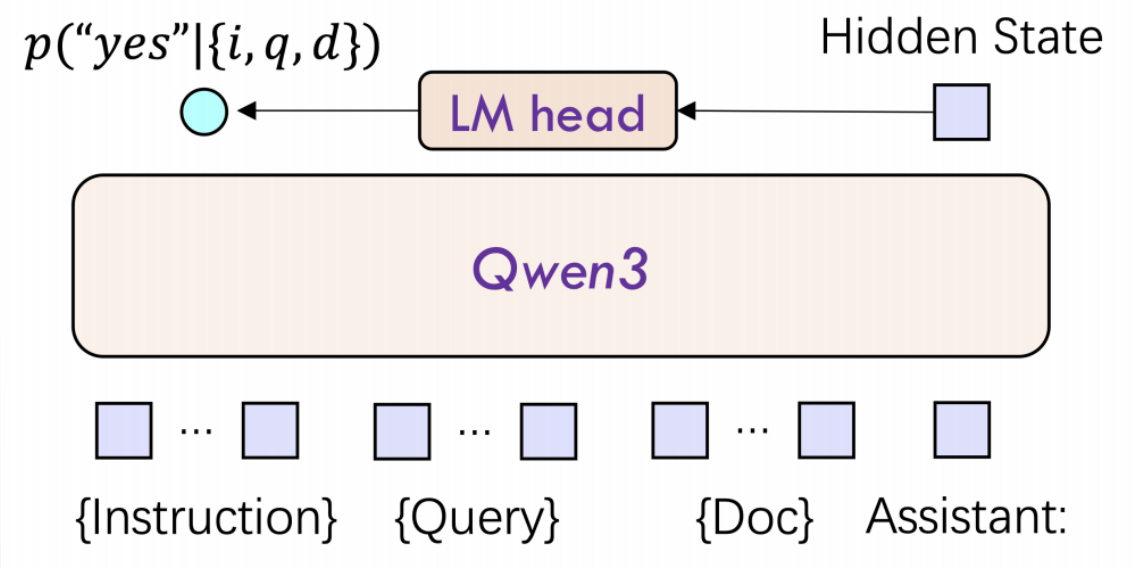
```
{
    'system': 'Judge whether the Document meets the requirements based on the Query and the Instruct provided. Note that the answer can only be "yes" or "no".',
    'instruction': '<Instruct>: Given a web search query, retrieve relevant passages that answer the query\n\n<Query>: ##query##\n\n<Document>: ##site_name##\n\n##summary_content##\n\n##site_content##',
    'input': '',
    'output': 'yes/no'
}
```

02 算法方案解析

Stage-1 SFT:

我们以Qwen3-Reranker-8B作为基座模型结合构建的数据进行监督指令微调来构建数据过滤模型,来识别潜在的错误标注样本.





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

02 算法方案解析

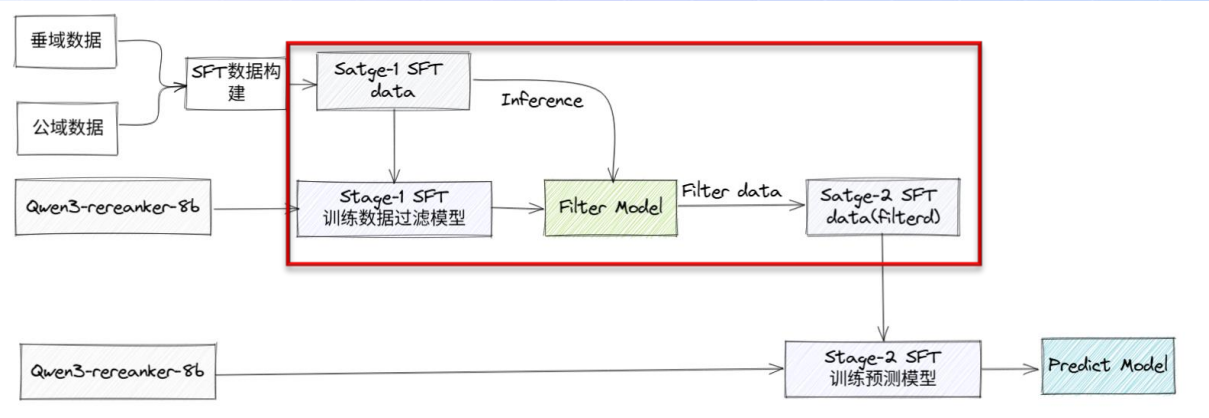
潜在错误样本检查:

基于构建的过滤模型结合通用的大语言模型来识别潜在的错误样本.

1. 利用过滤模型对原始训练集进行预测，识别两类可疑样本：
   - 模型预测概率在 **[0.25, 0.75]** 区间（低置信度）
   - 模型预测与原始标签不一致（疑似标注错误）

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

2. 调用更强 LLM（Qwen3-32B ）结合结构化 Prompt 对可疑样本进行自动校验，保留高质量 31K 样本.

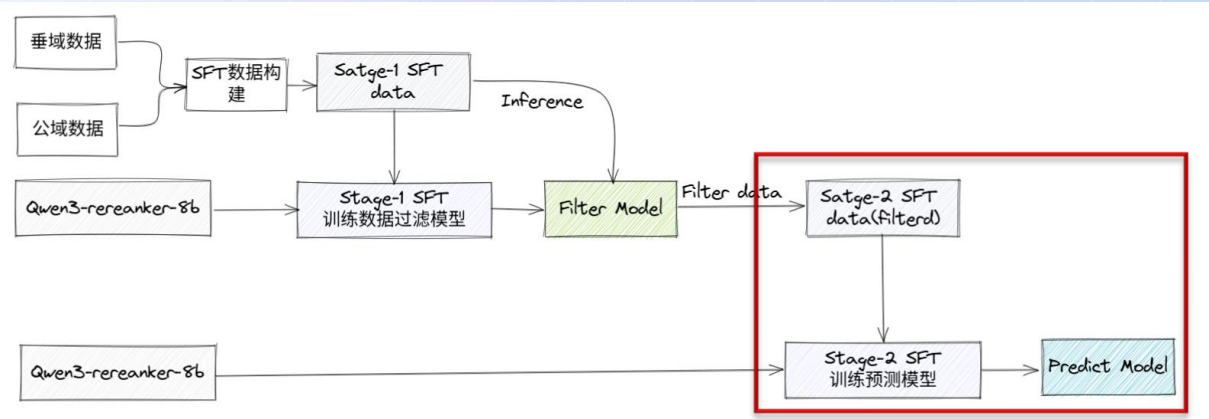



```
SYSTEM ="你是一个信息检索领域的专家，你需要咨询的阅读检索到的文档来判断当前文档是否可以回答用户的问题."
PROMPT ='''文档内容如下:
<document start>
<document_placeholder>
</document end>

用户问题是: <question><question_placeholder></question>

你的任务是根据提供的【文档内容】，判断是否能够充分回答用户的【问题】。请严格按照以下步骤进行分析，并以指定 JSON 格式输出结果。

### 分析步骤:
1. **理解问题**: 提取问题的关键词、意图和所需信息类型（如事实、定义、原因、步骤、数据等）。
2. **分析文档内容**: 查找文档中是否包含与问题相关的信息，判断是否:
   - 直接包含答案;
   - 可通过逻辑推理或归纳得出答案;
   - 缺失关键信息或完全无关。
3. **做出判断**: 基于证据决定是否可以回答。
4. **输出格式**: 请严格按照如下 JSON 格式输出，不要添加任何额外文本或解释:

{
  "analysis_content": "分析文档是否可以回答用户问题的详细分析步骤填充位置",
  "can_answer": "当前文档是否可以回答用户问题的判断结果,only yes or no"
}
'''
```

02 算法方案解析

Stage-2 SFT及模型预测:

基于过滤后的数据,我们基于Qwen3-Reranker-8B进行监督指令微调来构建预测模型并基于该模型对测试数据进行预测.



训练损失函数:

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

测试结果预测:

$$pred = \begin{cases} ture, & score(q,d) >= threshold \\ false, & score(q,d) < threshold \end{cases}$$

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

02 算法方案解析

实验设置:

训练设置:

基座模型：Qwen/Qwen3-Reranker-8B
微调方法：LoRA（rank=32, target=all） + SFT
框架：DeepSpeed
关键超参：
Epochs: 2
Batch size: 16
LR: 5e-5，cosine with min_lr
Warmup_ratio: 0.1

推理设置:
推理框架：VLLM
threshold：0.15

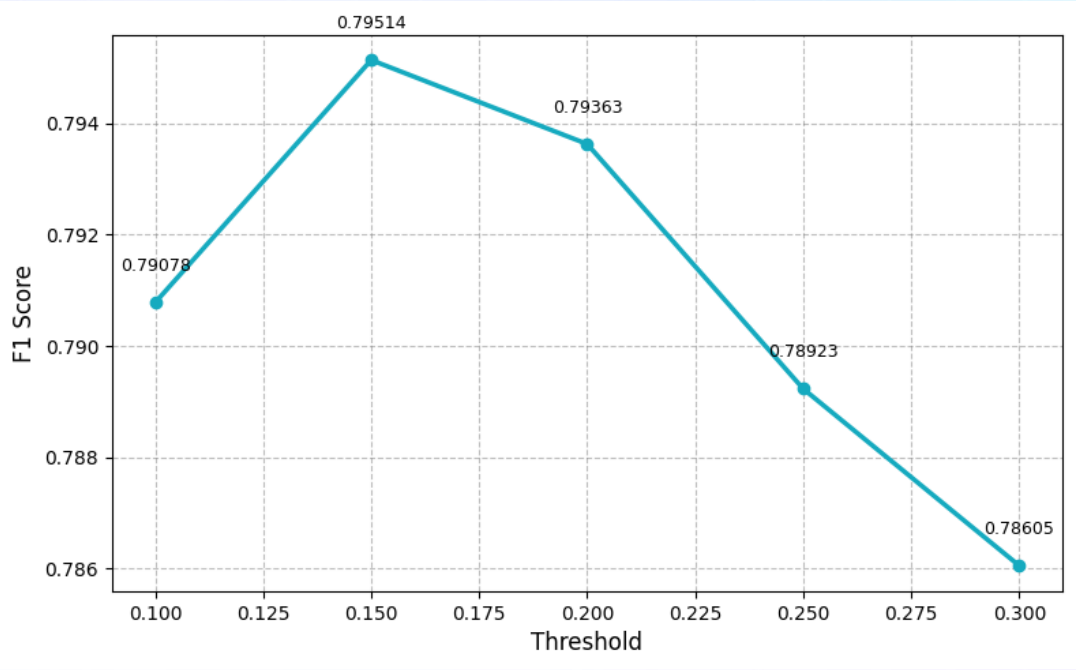
02 算法方案解析

实验结果:

Model	F1	ΔF1 (vs. Previous)
qwen3-0.6b-reranker	0.65000	-
qwen3-8b-reranker+校准分类边界	0.69962	+0.04962
qwen3-8b-reranker+Stage-1 SFT(未过滤数据)+校准分类边界	0.78957	+0.08995
qwen3-8b-reranker+Stage-2 SFT(过滤数据) +校准分类边界	0.79514	+0.00557

02 算法方案解析

实验结果:



Threshold 参数敏感性实验

Model	F1	ΔF1 (vs. Previous)
qwen3-8b-reranker+Stage-2 SFT(过滤数据)+校准分类边界	0.79514	-
qwen3-8b-reranker+Stage-2 SFT(过滤数据)+ 基于LLM合成数据(30k)+校准分类边界	0.78268	-0.01246

加入合成数据后性能下降的可能原因:

- 由于不能引入外部数据，合成数据只能基于提供的训练数据,数据多样性不足
- 受硬件限制,我们使用Qwen3-32B来进行数据合成,模型性能受限
- 合成的训练数据需要构建更严格的数据过滤Pipeline来进行数据过滤才能确保数据质量.

03 下阶段优化思路

1. **基于相关语料开展大规模数据合成**
   当前社区前沿成果（如 Qwen3-Embedding/Reranker[1]、Youtu-Embedding[2] 等）已验证，通过高质量的数据合成方案可显著提升模型性能。本赛题任务场景的相关语料（如[3]）可作为数据合成的核心原料，用于构建更具多样性的训练数据集.

2. **构建精细化的数据质量控制流水线**
   随着合成数据在训练中的广泛应用，Embedding 与 Reranker 模型的性能日益受制于数据质量。因此，亟需构建一套精细化的数据质量控制 pipeline，从源头对数据进行筛选、清洗与多样性增强，确保训练数据的准确性、覆盖度与鲁棒性，从而有效提升模型最终表现.

3. **平衡模型推理精度与速度**
   在实际部署场景中，模型推理效率直接影响系统响应延迟与资源消耗。当前使用的模型参数量大、推理延迟高，实际落地体验不佳。可以引入模型量化、知识蒸馏等轻量化策略，在核心任务指标损失可控的前提下，显著降低推理延迟，提升部署效率与实用性.

[1] Zhang Y, Li M, Long D, et al. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models[J]. arXiv preprint arXiv:2506.05176, 2025.
[2] Zhang B, Song Z, Chen C, et al. CoDiEmb: A Collaborative yet Distinct Framework for Unified Representation Learning in Information Retrieval and Semantic Textual Similarity[J]. arXiv preprint arXiv:2508.11442, 2025.
[3] Zhang, Yifan, et al. "Improving Chinese segmentation-free word embedding with unsupervised association measure." arXiv preprint arXiv:2007.02342 (2020).

交流讨论