FT-Data-Ranker 数据竞赛 7B 模型赛道技术报告

队伍：太棒了
队伍 ID：975209
队伍排名：2

1 数据处理

我们采用了 Baseline 提供的数据处理流程生成初版数据，根据对初版数据的分析，我们发现：1）一些数据组成如中文医疗数据仍然存在大量重复的情况；2）许多数据组成部分与测试集并没有关联。因此我们的方向在于调整数据去重策略与调整数据组成。具体来讲，我们的实验分为以下几个部分：

- 调整 simhash 的参数使其更加激进地去重；
- 搜索数据组成部分；
- 搜索随机种子，减小随机性带来的影响；
- 对数据比例进行初步探索；

默认的数据采样过程是直接进行均匀采样，这会导致数量较少的数据组成部分被注意到的权重较低。因此我们也尝试了对每一部分数据组成进行等概率采样。由于时间关系，我们并没有做对数据比例的进一步尝试。

2 实验结果

如表 1所示，我们测试了不同策略，并最终采用一阶段测试集分数最高的模型进行提交。由于可以看出本地验证集和在线测试集结果并不存在完全的正相关关系，因此我们以在线测试集结果为准进行调优。出于时间因素考虑，我们并未在数据比例上进行更大规模的搜索。

Data	Dev	Test
remove_zh_medical_en_convai2_eq_prob_seed_20000123	34.90	36.76
remove_zh_medical_en_convai2_seed_20000317	36.37	37.64
remove_zh_medical_en_convai2_seed_42	36.33	37.56
remove_zh_medical_en_convai2_seed_420	36.49	37.74
remove_zh_medical_en_convai2_seed_4200	36.21	37.46
remove_zh_medical_en_convai2_seed_20000123	35.73	**38.02**
remove_zh_medical_en_convai2_seed_0317	36.31	37.78
remove_zh_medical_en_convai2_simhash_12_hamming_10_seed_42	36.02	37.51
remove_zh_medical_en_convai2_simhash_12_hamming_10_seed_420	36.33	38.00
remove_zh_medical_en_convai2_simhash_12_hamming_10_seed_4200	35.91	
remove_zh_medical_en_convai2_simhash_12_hamming_10_seed_20000123	36.09	37.85
remove_zh_medical_en_convai2_simhash_12_hamming_10_seed_0317	35.87	

表 1: 实验结果。其中 eq_prob 表示对去除指定组成之后的剩余数据组成部分进行等概率采样。simhash_12_hamming_10 表示调整 simhash 的参数，使其去重行为更加激进。