**一、动机**

能否借鉴function call的工具调用任务形式，让LLM学会关注通过function description中的文本描述，对新的分类任务有一个较高的性能指标，

**二、优化实验前测试**

（都是在qwen3-8B上做的测试）

* **1、将prompt中描述都放入function description中后的测试效果：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Function desc | precision | recall | f1-score | support |
| gameTopic | 0.982206 | 0.948454 | 0.965035 | 291 |
| otherTopic | 0.90566 | 0.966443 | 0.935065 | 149 |
| accuracy | 0.954545 | 0.954545 | 0.954545 | 0.954545 |
| macro avg | 0.943933 | 0.957448 | 0.95005 | 440 |
| weighted avg | 0.956285 | 0.954545 | 0.954886 | 440 |

* **2、按照原来的，主要描述都写在prompt中，function description只是简单描述的测试效果：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| prompt | precision | recall | f1-score | support |
| gameTopic | 0.968966 | 0.965636 | 0.967298 | 291 |
| otherTopic | 0.933333 | 0.939597 | 0.936455 | 149 |
| accuracy | 0.956818 | 0.956818 | 0.956818 | 0.956818 |
| macro avg | 0.951149 | 0.952617 | 0.951876 | 440 |
| weighted avg | 0.956899 | 0.956818 | 0.956853 | 440 |

* **3、qwen3-8B上测试多分类fc效果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| bankTopic | 0.936893 | 0.639073 | 0.759843 | 302 |
| gameTopic | 0.964029 | 0.781341 | 0.863124 | 343 |
| insuranceTopic | 0.760563 | 0.924658 | 0.834621 | 292 |
| lawTopic | 0.958678 | 0.748387 | 0.84058 | 310 |
| mathTopic | 0.996587 | 0.986486 | 0.991511 | 296 |
| medicineTopic | 0.981013 | 0.984127 | 0.982567 | 315 |
| otherTopic | 0.158824 | 0.818182 | 0.26601 | 66 |
| telecomTopic | 0.954802 | 0.605735 | 0.741228 | 279 |
| travelTopic | 0.959184 | 0.989474 | 0.974093 | 190 |
| weatherTopic | 1 | 0.96997 | 0.984756 | 333 |
| accuracy | 0.84336 | 0.84336 | 0.84336 | 0.84336 |
| macro avg | 0.867057 | 0.844743 | 0.823833 | 2726 |
| weighted avg | 0.927734 | 0.84336 | 0.869468 | 2726 |

**三、实验对比**

两组实验：

1. sft + function desc 调优，ood效果（gameTopic在ood验证集，不在训练集）
2. sft + function desc 调优，id效果（gameTopic在id验证集，在训练集）

cutoff\_len: 2048

per\_device\_train\_batch\_size: 2

gradient\_accumulation\_steps: 8

learning\_rate: 5.0e-7

* **ood效果(qwen3\_lora\_sft\_ood\_less)**

{'loss': 0.5318, 'grad\_norm': 1.2736238241195679, 'learning\_rate': 7.342318745404034e-08, 'epoch': 0.78}

{'loss': 0.5376, 'grad\_norm': 1.5206336975097656, 'learning\_rate': 3.8254557187671374e-08, 'epoch': 0.84}

{'loss': 0.5195, 'grad\_norm': 1.5833200216293335, 'learning\_rate': 1.3856398044250844e-08, 'epoch': 0.91}

{'loss': 0.5224, 'grad\_norm': 1.5630214214324951, 'learning\_rate': 1.4697269024287196e-09, 'epoch': 0.97}

train\_loss = 0.8262

eval\_loss = 0.547

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| bankTopic | 0.918819 | 0.824503 | 0.86911 | 302 |
| gameTopic | 0.967972 | 0.793003 | 0.871795 | 343 |
| insuranceTopic | 0.961806 | 0.94863 | 0.955172 | 292 |
| lawTopic | 0.926471 | 0.812903 | 0.865979 | 310 |
| mathTopic | 1 | 0.989865 | 0.994907 | 296 |
| medicineTopic | 0.987382 | 0.993651 | 0.990506 | 315 |
| otherTopic | 0.184375 | 0.893939 | 0.305699 | 66 |
| telecomTopic | 0.97006 | 0.580645 | 0.726457 | 279 |
| travelTopic | 0.969072 | 0.989474 | 0.979167 | 190 |
| weatherTopic | 1 | 0.96997 | 0.984756 | 333 |
| accuracy | 0.876009 | 0.876009 | 0.876009 | 0.876009 |
| macro avg | 0.888596 | 0.879658 | 0.854355 | 2726 |
| weighted avg | 0.948098 | 0.876009 | 0.899555 | 2726 |

cutoff\_len: 2048

per\_device\_train\_batch\_size: 4

gradient\_accumulation\_steps: 8

learning\_rate: 1.0e-6

* **id效果(qwen3\_lora\_sft\_id)**

{'loss': 0.1603, 'grad\_norm': 2.6051905155181885, 'learning\_rate': 6.655307076412636e-08, 'epoch': 0.85}

{'loss': 0.1469, 'grad\_norm': 2.6380069255828857, 'learning\_rate': 3.299194563372604e-08, 'epoch': 0.9}

{'loss': 0.142, 'grad\_norm': 2.6623523235321045, 'learning\_rate': 1.0853984274007245e-08, 'epoch': 0.94}

{'loss': 0.1437, 'grad\_norm': 2.6331920623779297, 'learning\_rate': 6.806881768539052e-10, 'epoch': 0.99}

train\_loss = 0.5219

eval\_loss = 0.1478

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| bankTopic | 0.926471 | 0.834437 | 0.878049 | 302 |
| gameTopic | 0.96875 | 0.813411 | 0.884311 | 343 |
| insuranceTopic | 0.968531 | 0.94863 | 0.958478 | 292 |
| lawTopic | 0.927536 | 0.825806 | 0.87372 | 310 |
| mathTopic | 1 | 0.986486 | 0.993197 | 296 |
| medicineTopic | 0.987382 | 0.993651 | 0.990506 | 315 |
| otherTopic | 0.194079 | 0.893939 | 0.318919 | 66 |
| telecomTopic | 0.971264 | 0.605735 | 0.746137 | 279 |
| travelTopic | 0.969072 | 0.989474 | 0.979167 | 190 |
| weatherTopic | 1 | 0.96997 | 0.984756 | 333 |
| accuracy | 0.883346 | 0.883346 | 0.883346 | 0.883346 |
| macro avg | 0.891309 | 0.886154 | 0.860724 | 2726 |
| weighted avg | 0.950243 | 0.883346 | 0.905503 | 2726 |

* **ood效果(qwen3\_lora\_sft\_ood)**

{'loss': 0.3368, 'grad\_norm': 1.7622857093811035, 'learning\_rate': 8.132915035762695e-08, 'epoch': 0.84}

{'loss': 0.3294, 'grad\_norm': 1.847312092781067, 'learning\_rate': 3.8644851667255176e-08, 'epoch': 0.89}

{'loss': 0.3226, 'grad\_norm': 1.7606923580169678, 'learning\_rate': 1.1309214244396348e-08, 'epoch': 0.94}

{'loss': 0.3213, 'grad\_norm': 1.7954723834991455, 'learning\_rate': 2.316577429792299e-10, 'epoch': 0.99}

train\_loss = 0.6347

eval\_loss = 0.355

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| bankTopic | 0.925373 | 0.821192 | 0.870175 | 302 |
| gameTopic | 0.965517 | 0.816327 | 0.884676 | 343 |
| insuranceTopic | 0.958478 | 0.94863 | 0.953528 | 292 |
| lawTopic | 0.924188 | 0.825806 | 0.872232 | 310 |
| mathTopic | 1 | 0.986486 | 0.993197 | 296 |
| medicineTopic | 0.987382 | 0.993651 | 0.990506 | 315 |
| otherTopic | 0.196013 | 0.893939 | 0.321526 | 66 |
| telecomTopic | 0.971264 | 0.605735 | 0.746137 | 279 |
| travelTopic | 0.974093 | 0.989474 | 0.981723 | 190 |
| weatherTopic | 1 | 0.975976 | 0.987842 | 333 |
| accuracy | 0.882979 | 0.882979 | 0.882979 | 0.882979 |
| macro avg | 0.890231 | 0.885722 | 0.860154 | 2726 |
| weighted avg | 0.948654 | 0.882979 | 0.904595 | 2726 |

**四、实验数据分析**

整体precision、recall、f1-score变化

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| base | 0.927734 | 0.84336 | 0.869468 |
| ood（分类少） | 0.948098 | 0.876009 | 0.899555 |
| ood | 0.948654 | 0.882979 | 0.904595 |
| id | 0.950243 | 0.883346 | 0.905503 |