

OpenSeek 初赛技术报告

组名: jouw

一、选择训练集数据

对基线模型评测发现，模型在数学数据集 **GSM8K** 上的得分非常低，也即是模型的数学能力较弱。因此想法是优先提升模型推理能力，先补足短板，再去提升综合能力。在训练模型推理能力时，选择数据集的主要原则：（1）提升数学和代码语料的占比；（2）优先选择高质量数据，并保证 **cot** 的数据量；（3）因算力资源紧张，尽量能压缩训练的 **tokens** 量。

基线模型评测效果如下：

数据集	评测得分
HellaSwag	51.51
ARC(Average)	43.85
PIQA	72.91
MMLU(cloze)	32.7
CommonsenseQA	29.89
TriviaQA	15.28
Winograde	52.64
OpenBookQA	34.2
GSM8K(5-shot)	2.27
SIQA	42.12
CEval	31.18
CMMLU	30.63
平均得分	36.59833333

在综合考虑算力成本后，压缩了所选择的训练语料量，只筛选了 **14.88B** 的语料，其中数学和代码共 **7.95B**，占总语料的 **53.4%**。选择的 **cot** 数据 **4.4B**，占 **29.6%**。同时考虑到中文能力，故增加了语料 **zh_cc-high-loss0**。具体选择的数据集如下表所示：

Name	Tokens(B)	Tokens 占比
Nemotron-CC-high-actual-actual-high	1.14	7.661%
Nemotron-CC-high-synthetic-diverse_qa_pairs-high	0.56	3.763%
Nemotron-CC-high-synthetic-extract_knowledge-high	0.48	3.226%
Nemotron-CC-high-synthetic-knowledge_list-high	0.27	1.815%
arxiv	0.66	4.435%
books	0.48	3.226%
code-high	1.04	6.989%
cot_synthesis2_CC-high	0.39	2.621%
cot_synthesis2_code-high	0.47	3.159%

cot_synthesis2_math-high	1.35	9.073%
cot_synthesis2_math-mid	0.36	2.419%
cot_synthesis2_wiki-high	0.65	4.368%
cot_synthesis_math-high	0.52	3.495%
cot_synthesis_math-mid	0.66	4.435%
math-high	1.87	12.567%
math-mid	1.68	11.290%
wiki	0.43	2.890%
zh_cc-high-loss0	1.87	12.567%
Tokens 训练总量	14.88	100%

二、模型训练

1. 训练脚本

由于使用是北京超级云计算中心的算力，为了适配该平台提交训练任务，将训练脚本最终修改为以下脚本：

<https://github.com/jouw/OpenSeek/blob/competition-jouw/openseek/competition/pz/jouw/preliminary-round/submit-training-job.sh>

主要修改如下：

(1) 对 learning rate 设置为 1e-5, 前 50 步 warmup 再进行余弦 decay, 并增加 weight decay

```
--lr 0.00001 --min-lr 0.0000001 --lr-warmup-iters 50 --lr-warmup-samples 0 --lr-decay-style cosine --weight-decay 0.01
```

说明：原始的 lr 较大导致模型初始训练 loss 不容易收敛，调整后模型收敛良好

(3) 设置总训练 iters

```
--train-iters 3620
```

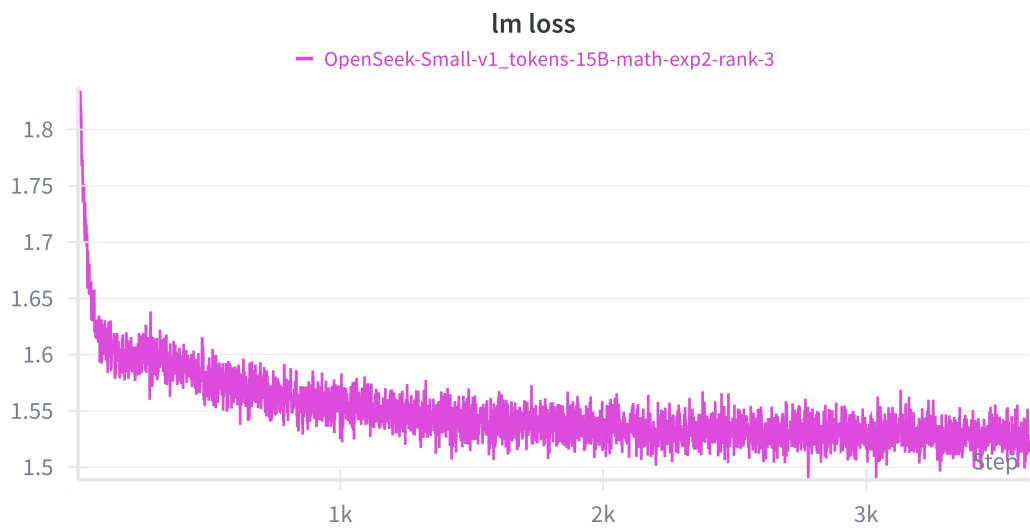
(4) 将语料按比例切分为训练集，验证集和测试集

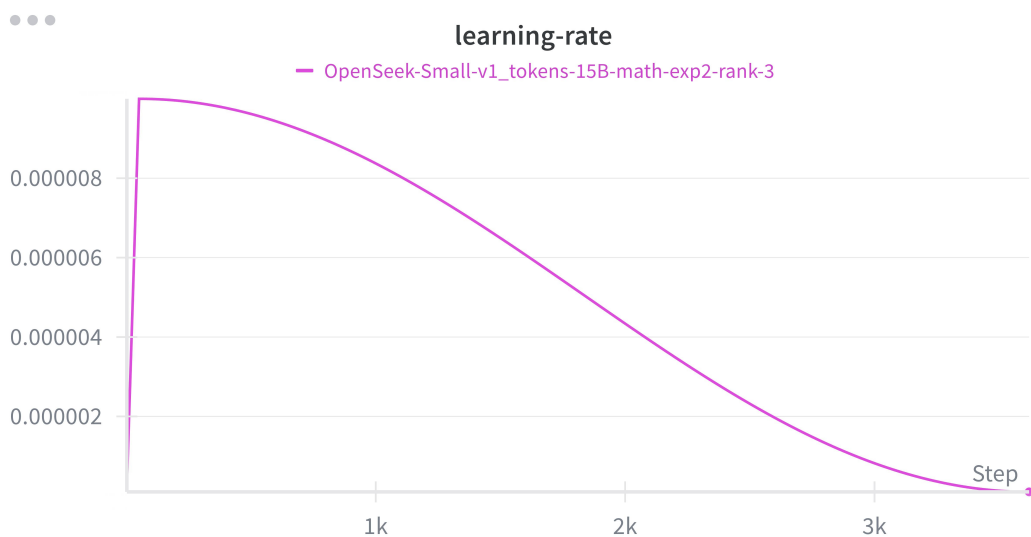
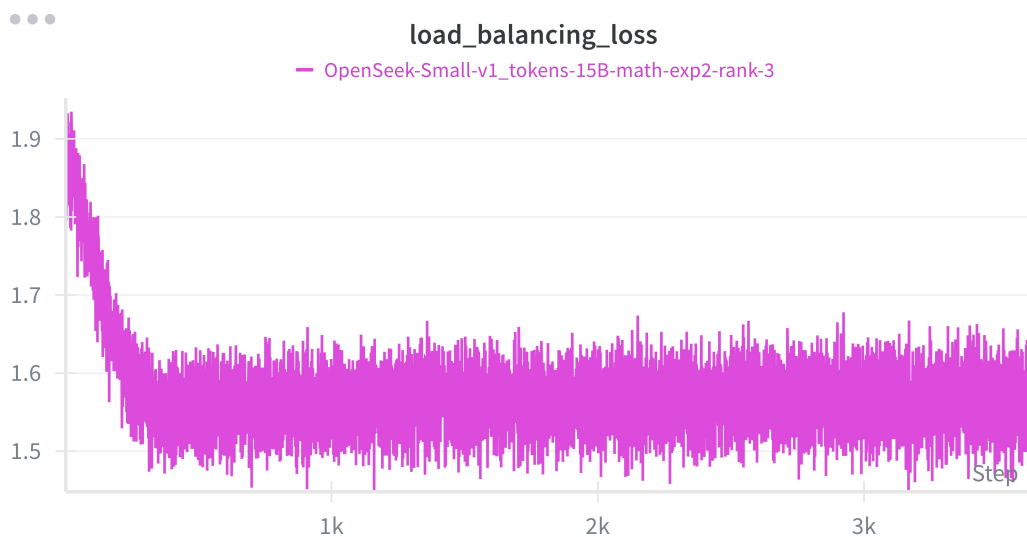
```
--split 998,1,1
```

(5) 4 卡 A100 训练

```
export CUDA_VISIBLE_DEVICES=0,1,2,3
--nproc_per_node 4
```

2. 训练日志





详细请参考 wandb 报表: <https://api.wandb.ai/links/oceanplus/zt033huk>

三、评测效果

最终模型评测得分 **37.31**，具体评测数值如下：

数据集	Baseline	iter_000072 0	iter_0001 440	iter_00021 60	iter_00028 80	iter_00036 00
HellaSwag	51.51	52.73	52.77	52.71	52.65	52.76
ARC(Average)	43.85	46.58	46.77	46.97	47.41	47.39
PIQA	72.91	72.91	73.39	73.39	73.39	73.78
MMLU(cloze)	32.7	33.03	33.26	33.11	32.98	33.17
CommonsenseQA	29.89	31.7	31.2	31.61	30.79	31.37
TriviaQA	15.28	14.72	16.16	15.96	16.15	16.16

Winograde	52.64	52.25	52.09	52.01	52.09	51.93
OpenBookQA	34.2	34	35.2	35	35.2	35.4
GSM8K(5-shot)	2.27	2.58	2.65	2.58	2.88	2.81
SIQA	42.12	42.43	42.12	42.02	41.81	41.71
CEval	31.18	29.85	28.98	30.41	29.92	30.17
CMMLU	30.63	30.73	30.76	30.85	31.03	31.07
最终得分	36.59833 333	36.9591666 7	37.1125	37.218333 33	37.191666 67	37.31