OpenSeek 共创大模型挑战赛初赛技术报告

1. 数据选取策略与设计理念

数据是模型能力的基石。本次优化的核心思路是，通过构建一个高质量、目标导向的数据集，来提升 CPT Base 模型的综合表现。

1.1 总体原则

我从总计 100B 的数据池中，筛选出 30B 数据用于本次训练。筛选的核心原则是质量优先，即优先选择被标记为 high 质量的数据类别，以确保模型学习的效率与效果。

1.2. 数据集构成与权重分配

为了实现目标，我融合了多个来源的数据，并为其分配了不同的权重，具体构成如表 1 和图 1 所示：

表 1 数据集构成

Category	Weight	Category	Weight
Nemotron-CC-high-actual-actual-high	1.1068	code-high	1.0102
Nemotron-CC-high-actual-actual-low	0.3577	cot_synthesis2_CC-high	0.3755
Nemotron-CC-high-actual-actual-mid	0.7775	cot_synthesis2_OpenSource-high	0.2573
Nemotron-CC-high-synthetic-distill-high	0.2859	cot_synthesis2_arxiv-high	6.0237
Nemotron-CC-high-synthetic-distill-low	0.1672	cot_synthesis2_code-high	0.4598
Nemotron-CC-high-synthetic-distill-mid	0.2339	cot_synthesis2_math-high	1.3135
Nemotron-CC-high-synthetic-diverse_qa_pairs-high	0.5397	cot_synthesis2_wiki-high	0.6314
Nemotron-CC-high-synthetic-diverse_qa_pairs-low	0.4064	cot_synthesis_CC-high	0.2225
Nemotron-CC-high-synthetic-diverse_qa_pairs-mid	0.5005	cot_synthesis_OpenSource-high	0.4081
Nemotron-CC-high-synthetic-extract_knowledge-high	0.4616	cot_synthesis_arxiv-high	5.68
Nemotron-CC-high-synthetic-extract_knowledge-low	0.067	cot_synthesis_code-high	0.7663
Nemotron-CC-high-synthetic-extract_knowledge-mid	0.3429	cot_synthesis_math-high	0.5074
Nemotron-CC-high-synthetic-knowledge_list-high	0.261	cot_synthesis_wiki-high	0.4
Nemotron-CC-high-synthetic-knowledge_list-low	0.1824	math-high	1.8165
Nemotron-CC-high-synthetic-knowledge_list-mid	0.2313	zh_cc-high-loss0	0.687
Nemotron-CC-high-synthetic-wrap_medium-high	0.8237	zh_cc-high-loss1	0.9776
Nemotron-CC-high-synthetic-wrap_medium-low	0.2866	zh_cc-high-loss2	0.3725
Nemotron-CC-high-synthetic-wrap_medium-mid	0.667	arxiv	0.1
books	0.29		

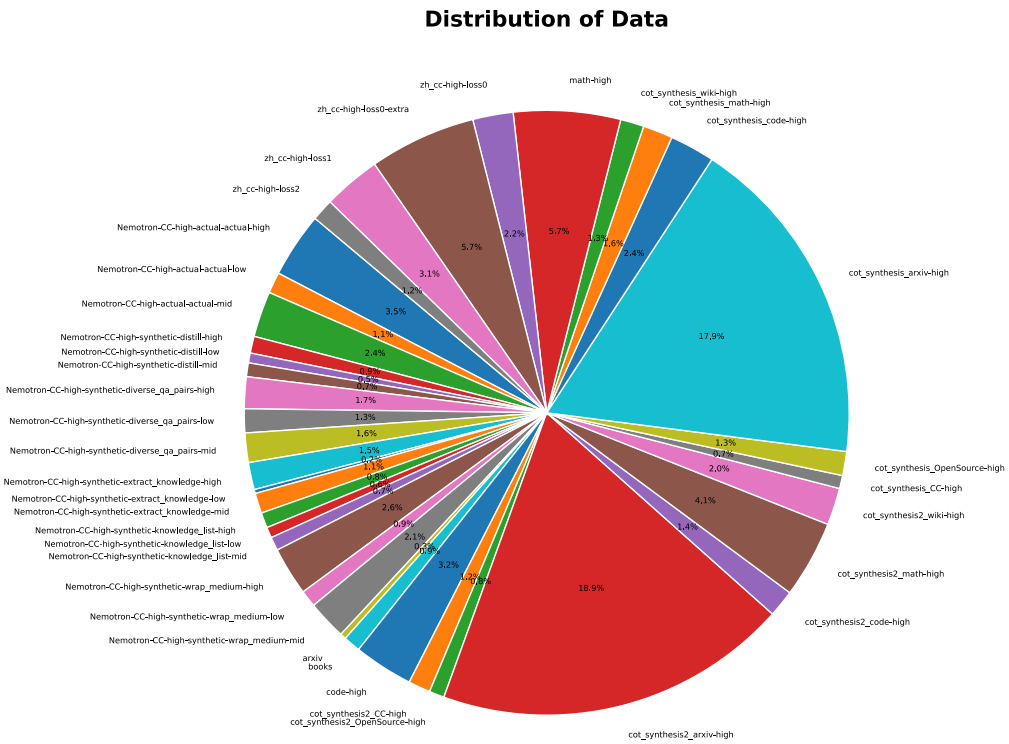

图 1 各类数据占比

1.3 数据分布与设计理念

为了更清晰地展示我的设计意图，我将上述数据划分为四个主要类别，如表 2 所示：

表 2 数据大类别占比

Category	Total Weight
Chain-of-Thought Synthesis (CoT Synthesis)	17.0455
General Web Text (Nemotron-CC)	7.6991
Domain-Specific Data	3.2167
Chinese Web Text (zh_cc)	2.0371

在训练语料的构成上，不同类别的数据权重分布体现了明确的设计取向。首先，Chain-of-Thought Synthesis (17.05) 占据了整体数据的最大比重，表明我希望训练重点在于强化逐步推理与复杂问题分解能力。其次，General Web Text (7.70) 提供了丰富的通用语言与常识支撑，为模型的语言理解与生成打下基础。在此之外，Domain-Specific Data (3.22) 作为重要补充，涵盖学术、代码、数学等专业领域，旨在提升模型在特定任务中的表现。最后，Chinese Web Text (2.04) 虽占比较低，但保证了模型在中文语境下的基本适应能力。总体来看，这一数据构成策略体现了“推理能力优先，通用能力支撑，专业能力补充，中文能力覆盖”的训练思路。

2. 模型训练配置

在训练阶段，我对超参数只进行了微调，具体如下：

学习率策略 (Learning Rate Scheduler): 采用 Cosine Decay 策略，使学习率平滑下降。

最大学习率 (Initial Learning Rate): 1e-5
最小学习率 (Minimum Learning Rate): 1e-6
微批次大小 (Micro Batch Size): 4

3. 实验结果

在训练 7200steps 后，在测试集上分数达到 37.2，初赛排名第三位。

4. 反思与展望

4.1 工作反思

1.数据质量问题： 在整理报告的过程中，我复盘发现 Nemotron-CC-high 分类下仍然选取了部分 low 和 mid 质量的样本。这一疏忽可能在一定程度上影响了数据的整体纯度，对模型性能造成了潜在的负面影响。

2.训练调优不足： 本次实验在超参数调优方面投入有限，当前配置仅为一组基线设置。更精细化的调优工作（如调整学习率、优化器参数等）或许能带来进一步的性能提升。

4.2 未来工作展望

基于本次的实践与反思，我规划了两个主要的优化方向：

1.调整损失函数以缓解“灾难性遗忘”： CPT 类模型在微调时常存在“灾难性遗忘”问题。借鉴以往自身大模型实践经验，我计划在损失函数层面进行优化。具体而言，可以在主损失函数（Cross-Entropy）的基础上，引入一个 reference model (即 base 模型)，并增加一项 KL 散度正则化约束。该约束用于惩罚训练模型 (train model) 与参考模型在输出分布上的巨大差异，通过这种方式在学习新知识与保留旧能力之间取得平衡。

2.引入质量评分模型实现数据分层采样： 如果严格限定只使用 high 质量数据，可能无法满足 30B 的数据量需求。为此，我提出一个更精细化的数据分层策略：使用一个独立的“文本质量评分模型” 对所有 mid 和 low 质量的数据进行打分。然后，根据分数进行加权采样或设定阈值进行过滤，从而在保证数据量的同时，最大化地提升中低质量数据的可用性，实现权重的动态调整。