

超越杯算法赛道--初赛技术报告

参赛单位：北京城建智控科技股份有限公司
团队名称：UCI001
提交账号：liximeng0824

一、数据分析

对官方提供 100B 数据集的子集进行数据抽样展示和分析。

专业上可分为：百科、代码、通识知识、论文、数学、科学等；

形式上可分为：预训练 PT、问题分析 CoT

二、数据配比

原则 1：选择形式为问题分析类的数据集，只去除掉“代码问答+生成”（如 code-xxx、cot_synthesis_code-xxx cot_synthesis_code-high）。

问题分析 CoT 类数据包含解决问题步骤拆解、逻辑推理，是真正有效的 CoT 数据；

除掉“代码问答+生成”的主要原因是其回答直接使用大片纯代码段，不包含任何场景描述和问题分析过程，团队认为在持续预训练阶段采用该类数据无法提升通用能力。

原则 2：选择高质量预训练 PT 形式数据，尽量包含各专业（不超过 30B），主要包含以下：

论文及延伸思考：arxiv、cot_synthesis2_arxiv-high

科学常识类 COT：cot_synthesis2_CC-high

数学：cot_synthesis2_math-high

百科：cot_synthesis2_wiki-high

通识知识：Nemotron-CC-high-synthetic-diverse_qa_pairs-high、

Nemotron-CC-high-synthetic-extract_knowledge-high

训练数据子集选择时全部使用 high 进行，没有再进一步分析 high/mid/low 的影响。

原则 3：由于初始模型的中文能力偏弱，需要增加中文语料 zh_cc-high-loss0，取 loss0（据工作人员反馈为困惑度最低的含义）

三、训练方案

记录团队最优的两套方案，全部为数据策略，没有对模型结构进行优化

方案 1（保底）：单阶段训练，保守策略

数据集为 22.57B，训练配置中 train_samples 为 6144000（25B），学习率 3e-5

评测指标为 37.2875，请见配置 train_deepseek_v3_1_4b_test6.yaml

方案 2：两阶段训练

1、第一阶段：采用 PT + 问题分析 CoT 混合数据，即方案 1。

2、第二阶段：基于第一阶段训练 checkpoint，采用问题分析 CoT 进行训练。

数据集为 4.64B，训练配置中 train_samples 为 1228800（5B），学习率 1e-5

评测指标为 37.5733，请见配置 train_deepseek_v3_1_4b.yaml