使用的基模型：

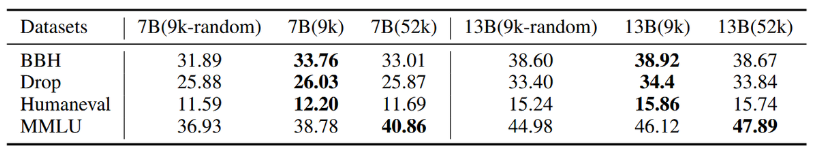
LLama2-7b, LLama2-13b, mistral-7b,

使用的数据集：

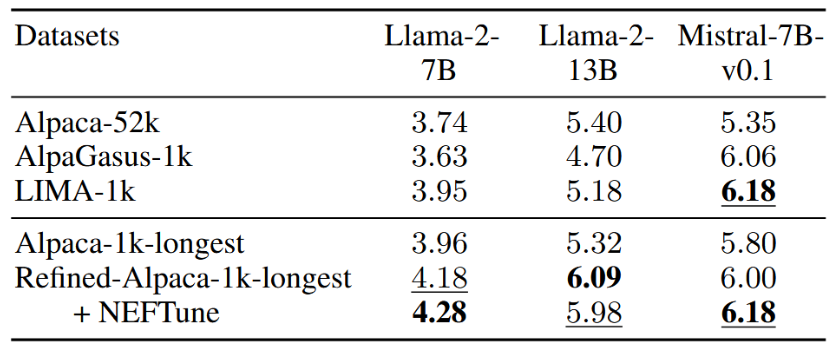
52k Alpaca dataset, Evol-Instruct-70k

对标论文的baseline&benchmark:

(1)AlpaGasus: Training A Better Alpaca with Fewer Data



(2)Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning-2024.6（ICML-A接收）



打标签参考论文：

INSTAG: INSTRUCTION TAGGING FOR ANALYZING SUPERVISED FINE-TUNING OF LARGE LANGUAGE MODELS

目前要做的实验：

1. 聚类数量效果探究：1%，5%，10%，···
2. 打标签的方法路径研究