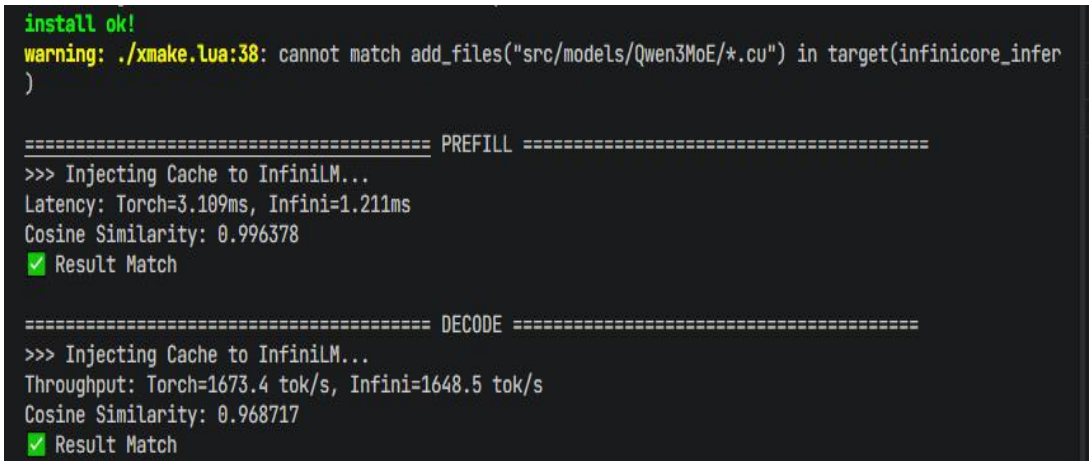
InfiniTensor T2-1-1

队伍：newbie!　成员 刘轩睿，张骏颉



实现平台：NVIDIA
上述为 NVIDIA 平台的测试截图
Prefill TTFT 约为 torch 的 40%，Decode 的 throughput 与 torch 相似
使用 Cosine Similarity 分析输出正确性，Prefill > 0.99，decode>0.96 考虑到 decode 阶段 100 轮推理可能将精度误差放大，可以接受

实现中算子全部为 InfiniOps，部分使用 InfiniRT，部分使用 cudaRuntime
采用静态预分配 KVCache，提前分配一个 MaxSeqlen 的 KVCache

未完全实现部分：
有部分方法未使用 Infinirt，例如 kvcache 部分可以考虑替换 rearrange

Padding 部分需要用 cudaMemset 好像没找到有 infiniMemset 的方法实现...