Experiment of Intel Extension for Transformers

## 1 Intel技术介绍

英特尔 Extension for Transformers 是英特尔推出的一个创新工具包，可基于英特尔架构平台，尤其是第四代英特尔至强可扩展处理器（代号 Sapphire Rapids，SPR）显著加速基于 Transformer 的大语言模型 (Large Language Model, LLM)。其主要特性包括：

通过扩展 Hugging Face transformers API 和利用英特尔® Neural Compressor，为用户提供无缝的模型压缩体验；

提供采用低位量化内核（NeurIPS 2023：在 CPU 上实现高效 LLM 推理）的 LLM 推理运行时，支持 Falcon、 LLaMA、MPT、 Llama2、 BLOOM、 OPT、 ChatGLM2、GPT-J-6B、Baichuan-13B-Base、Baichuan2-13B-Base、Qwen-7B、Qwen-14B 和 Dolly-v2-3B 等常见的 LLM；

先进的压缩感知运行时（NeurIPS 2022：在 CPU 上实现快速蒸馏 和 QuaLA-MiniLM：量化长度自适应 MiniLM；NeurIPS 2021：一次剪枝，一劳永逸：对预训练语言模型进行稀疏/剪枝）。

## 2 实验过程及结果

1. 进入ModelScope，右上角点击完成新用户的注册



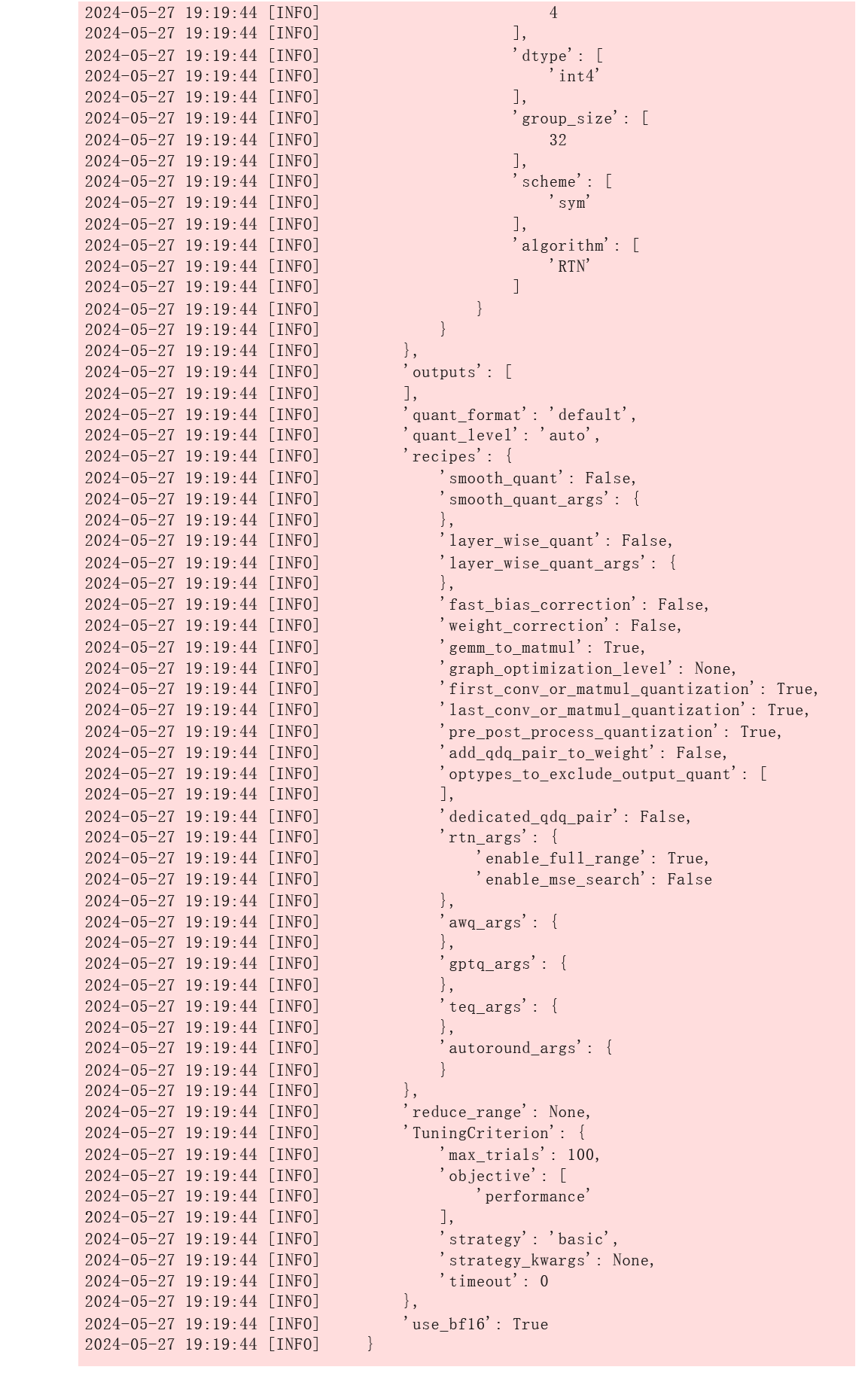
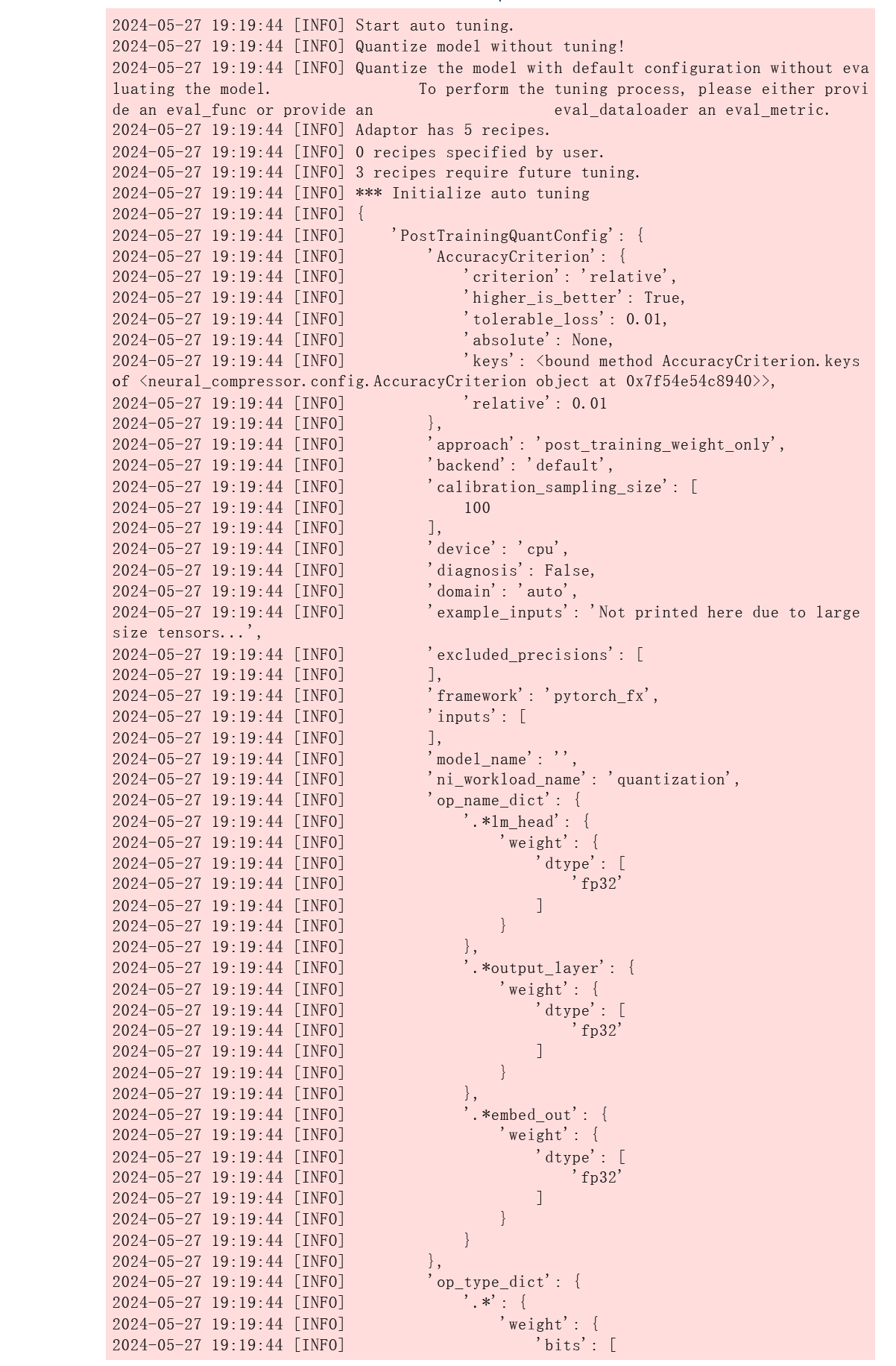
1. 注册完成后，登录ModelScope，进入首页，绑定阿里云账号，获得免费的配套云计算资源，启动CPU服务器



1. 打开terminal，进入环境目录/opt/conda/envs, 新建文件夹itrex，然后wget得到文件并解压到itrex文件目录，激活环境并安装对应的kernel，得到itrex kernel



1. 基于itrex kernel 新建 notebook，并进行如下notebook中的操作，并且获得图中的实验结果



## 3 个人心得

在Intel Extension for Transformers的实验中，在使用创新工具的情况下，Transformer-based LLM的inference速度显著提升了，这个工具包含了一个名为LLM Runtime的运行时环境，专门为intel处理器进行了优化，利用Intel的各种AI加速技术，如AMX、VNNI等，就可以实现更快速高效的LLM Inference。

同时，LLM Runtime还提供了多种量化选择，更好地适应应用场景。

在实验过程中，这个工具包极大提高了我的工作效率，在处理LLM时加速了我的推理任务，减少了资源消耗，能更好地应对LLM在各种应用场景中的落地，提高实际体验。

总的来说，Intel Extension for Transformers提供了一套高效、便利的工具，为我在Intel架构平台上开发和部署Transformer-based LLM提供了很大的帮助。我相信在未来的工作中，这个工具包会发挥很大的作用，帮助我更好地处理和应对各种复杂的NLP任务。