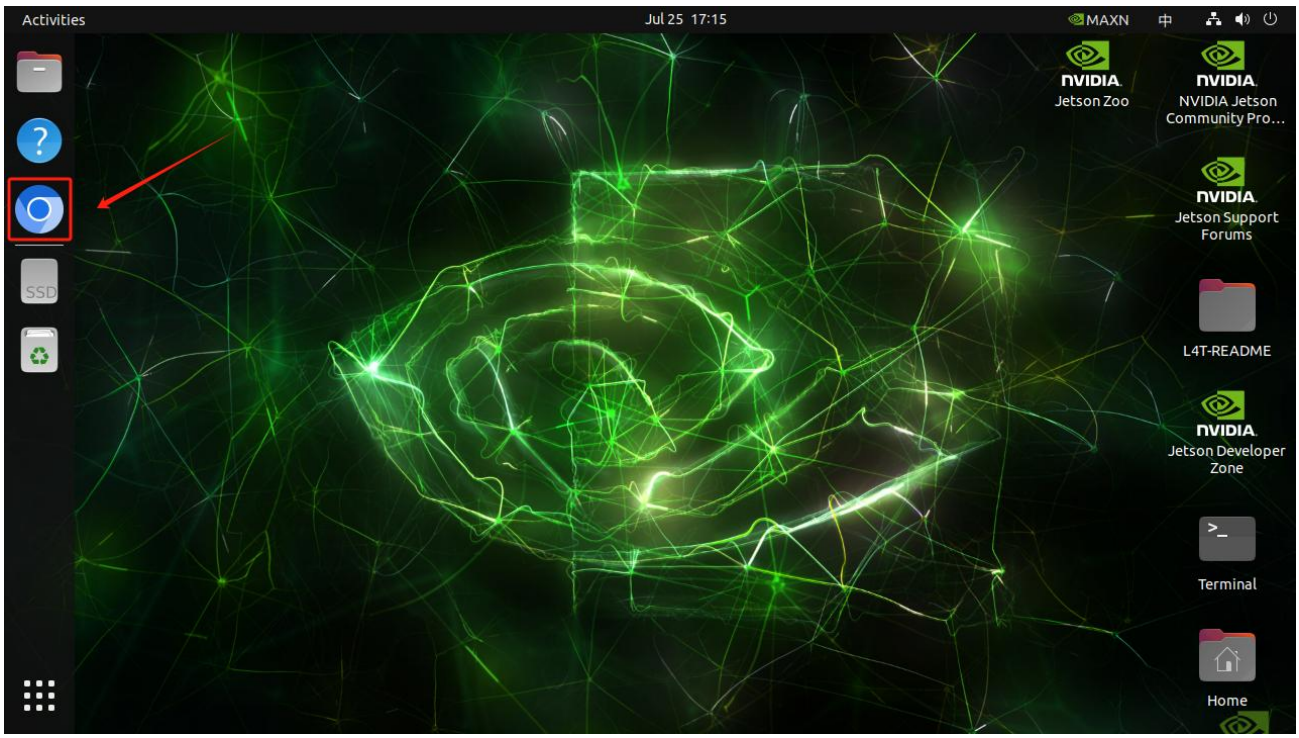
图为大模型一体机使用手册

1. 设备上直接访问使用

设备接上电源，使用 HDMI 线连接显示器与键鼠设备

用户名：nvidia 密码：nvidia，输入密码后，进入系统。

点击谷歌浏览器

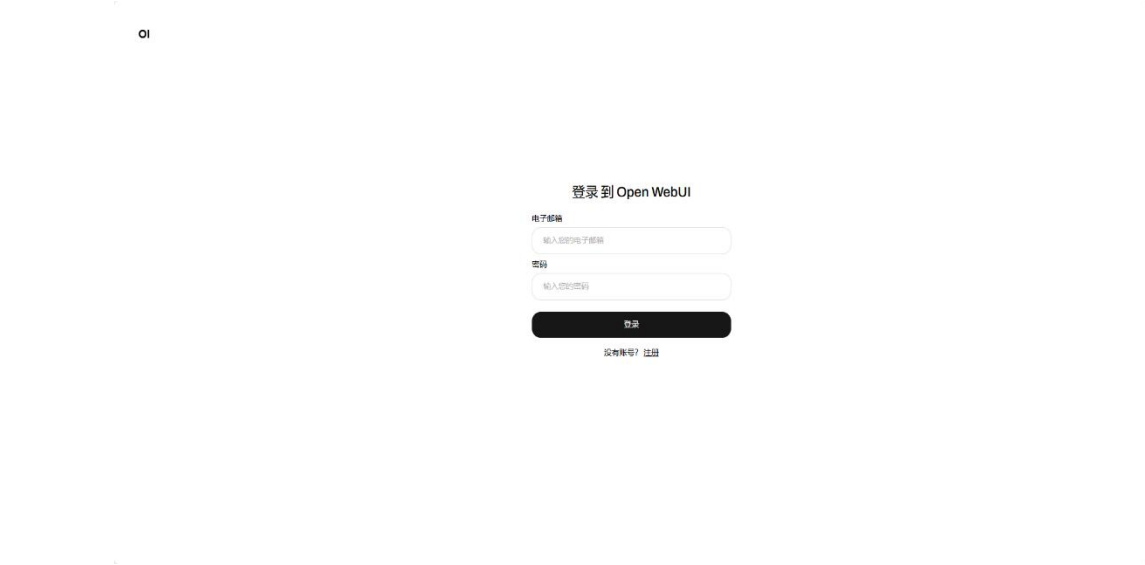


打开 谷歌浏览器 会自动跳转至下方界面

如没有跳转 可通过 http://0.0.0.0:8080/ 进行访问

初始管理员账号密码 账号：root@163.com 密码：root

可在下方注册新账号（账号信息在本地存储，不会上传云端，注册邮箱的格式正确即可通过验证，不是真实邮箱也可以哦）



登录成功后可进入到该界面

可在下方通过文字或语音与内置大模型进行交互(语音服务需要接入麦克风到设备上)

在左上角可以更换模型

2. 局域网内访问使用

首先查看设备的 IP 地址：

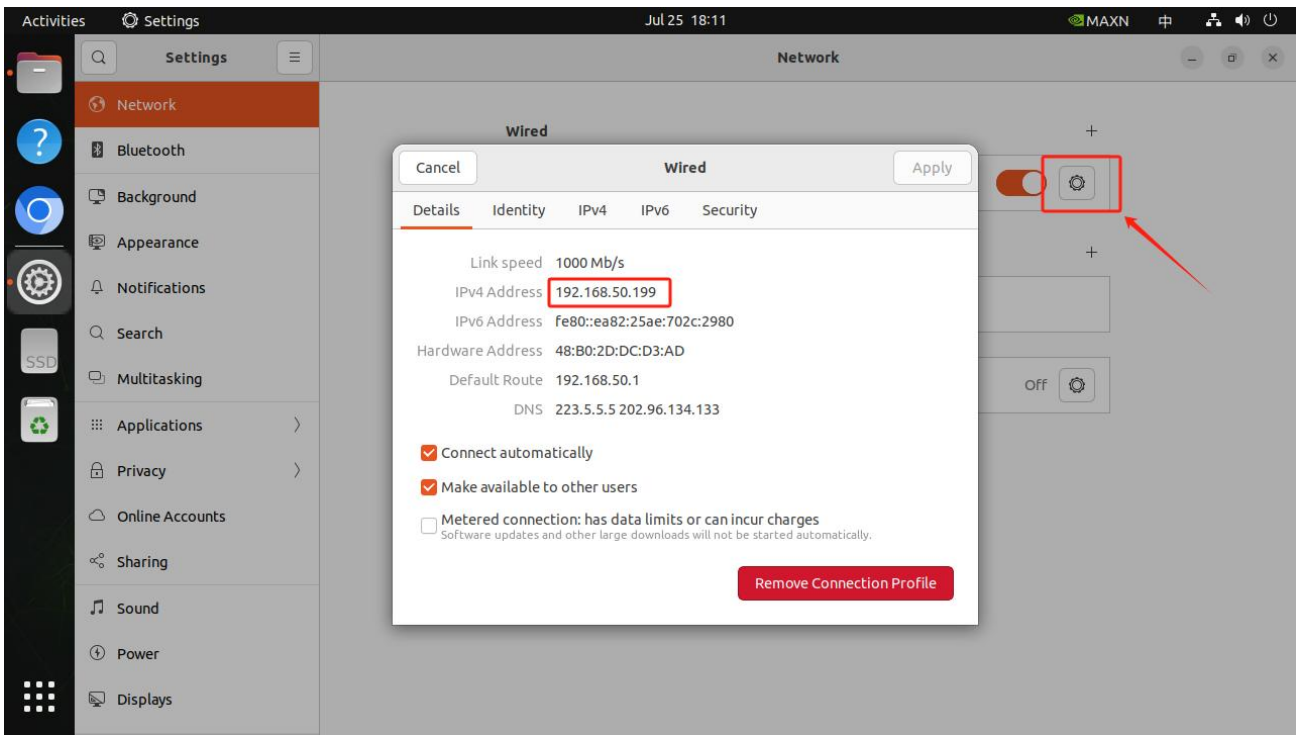
左侧打开设备终端 Terminal，或者快捷方式 **Ctrl + Alt + T** 打开终端。输入 **ifconfig eth0** 或 **ifconfig eth1** 可以查看到 IP 地址为：192.168.50.199

（eth0 或 eth1，取决于插入不同的网口，可以两个都试试）

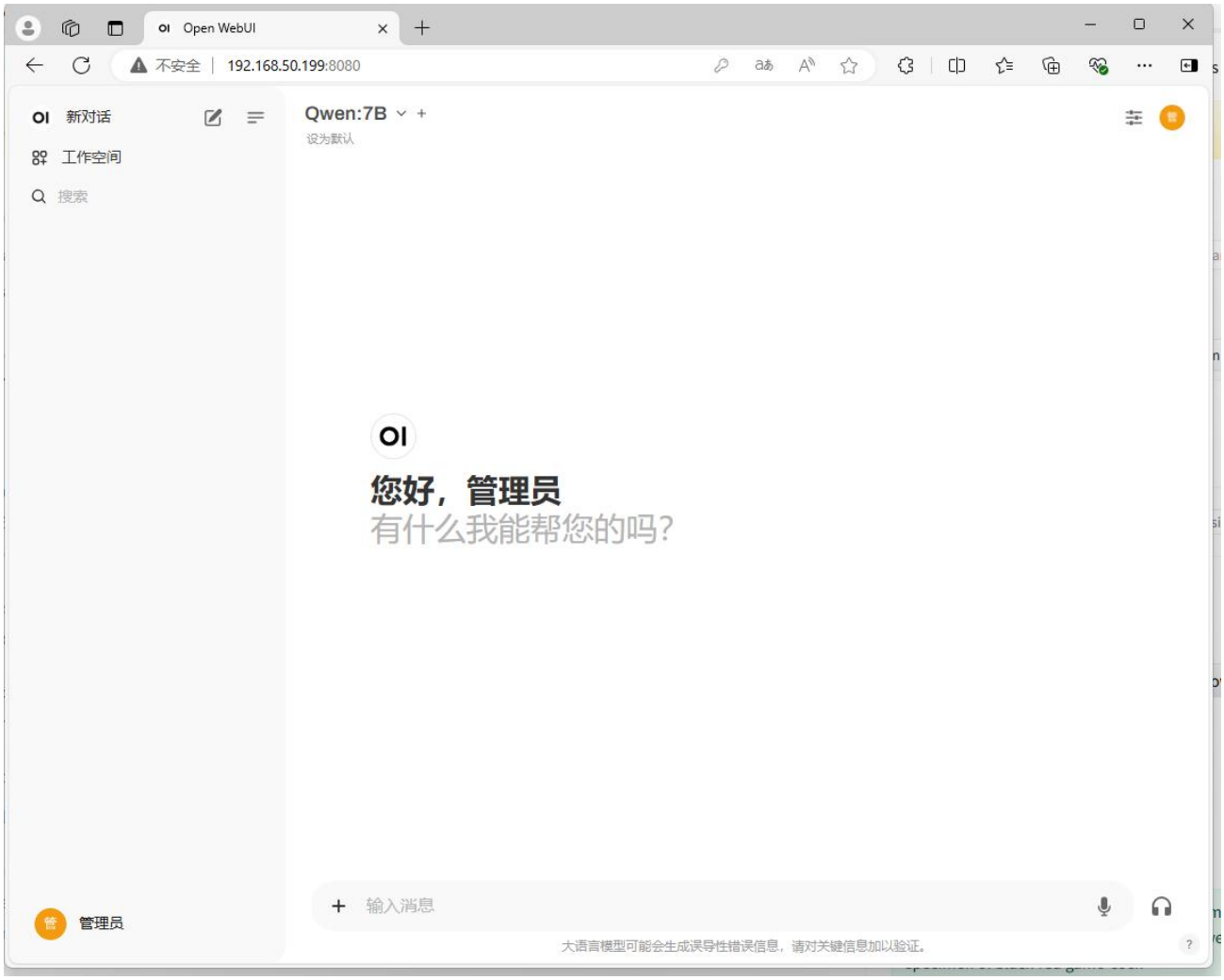
或者在网络设置中也可以查看 IP 地址：打开设置，点击网络连接的设置按钮，可以查看到

```
(base) nvidia@tegra-ubuntu:~$ ifconfig eth0
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1466
        inet 192.168.50.199  netmask 255.255.255.0  broadcast 192.168.50.255
        inet6 fe80::ea82:25ae:702c:2980  prefixlen 64  scopeid 0x20<link>
        ether 48:b0:2d:dc:d3:ad  txqueuelen 1000  (Ethernet)
        RX packets 23636  bytes 2169731 (2.1 MB)
        RX errors 0  dropped 1277  overruns 0  frame 0
        TX packets 3852  bytes 3124110 (3.1 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

IPv4 的地址 为：192.168.50.199



在局域网内的其它电脑或手机打开浏览器，输入：192.168.50.199:8080 稍等片刻，登录后即可与大语言模型对话



3. 内置模型介绍

通义千问（Qwen），是阿里云精心打造的一款大型语言模型，属于通义千问大模型系列包含多种参数规模，在本设备中部署了 70 亿参数，140 亿参数两种规格。Qwen 基于先进的 Transformer 架构，经过超大规模预训练数据的学习，具备了强大的语言处理能力。这些预训练数据种类丰富，覆盖网络文本、专业书籍、代码等多种类型。

模型细节（Model）

Qwen-7B 模型具体参数如下：

- 层数（n_layers）：32
- 注意力头数（n_heads）：32
- 模型维度（d_model）：4096
- 词汇表大小（vocab size）：151851
- 序列长度（sequence length）：8192

Qwen-14B 模型具体参数如下：

- 层数（n_layers）：40
- 注意力头数（n_heads）：40
- 模型维度（d_model）：5120
- 词汇表大小（vocab size）：151851
- 序列长度（sequence length）：2048

Qwen 模型架构

在模型架构方面，采用了目前流行的技术，包括 RoPE 相对位置编码、SwiGLU 激活函数和 RMSNorm 标准化处理（可选安装 flash-attention 加速）。在分词器方面，Qwen 使用了超过 15 万 token 的词表，该词表在 GPT-4 使用的 BPE 词表 cl100k_base 基础上进行了优化，对中文、多语言更加友好，提高了对中、英、代码数据的高效编解码能力。

Qwen 模型的训练方法

Qwen 模型的训练方法主要包括分布式并行加速、算法模型架构和内存计算优化三个方面。首先，通过数据并行、模型并行、流水线并行和张量并行等四种并行方式，Qwen 模型能够充分利用多台机器的计算资源，实现高效的训练。其次，Qwen 模型采用了 Transformer 网络模型结构和专家混合模型 MoE 等创新算法模型结构，提高了模型的训练速度和精度。最后，通过内存优化技术如激活重计算、内存高效的优化器和模型压缩，以及计算优化技术如混合精度

训练、算子融合和梯度累加等，Qwen 模型在训练过程中能够更有效地利用内存和计算资源。

Qwen 模型的应用场景

Qwen 模型凭借其高效的性能和广泛的应用领域，已经在自然语言处理、计算机视觉、语音识别等多个领域取得了显著的成果。例如，在自然语言处理领域，Qwen 模型可以应用于文本分类、情感分析、机器翻译等任务；在计算机视觉领域，Qwen 模型可以用于图像识别、目标检测等任务；在语音识别领域，Qwen 模型可以提高语音识别的准确率和速度。