1、下载和安装cherry-studio

下载页面：https://docs.cherry-ai.com/cherrystudio/download

建议直接安装在笔记本电脑

2、模型设置

1）左下角，设置->模型服务->添加：

提供商名称：取一个名字如zepdi

提供商类型：OpenAI

API密钥：empty

API地址(注意最后那个"/"不能少)：http://powerai.cc:28001/v1/

2）模型->添加：

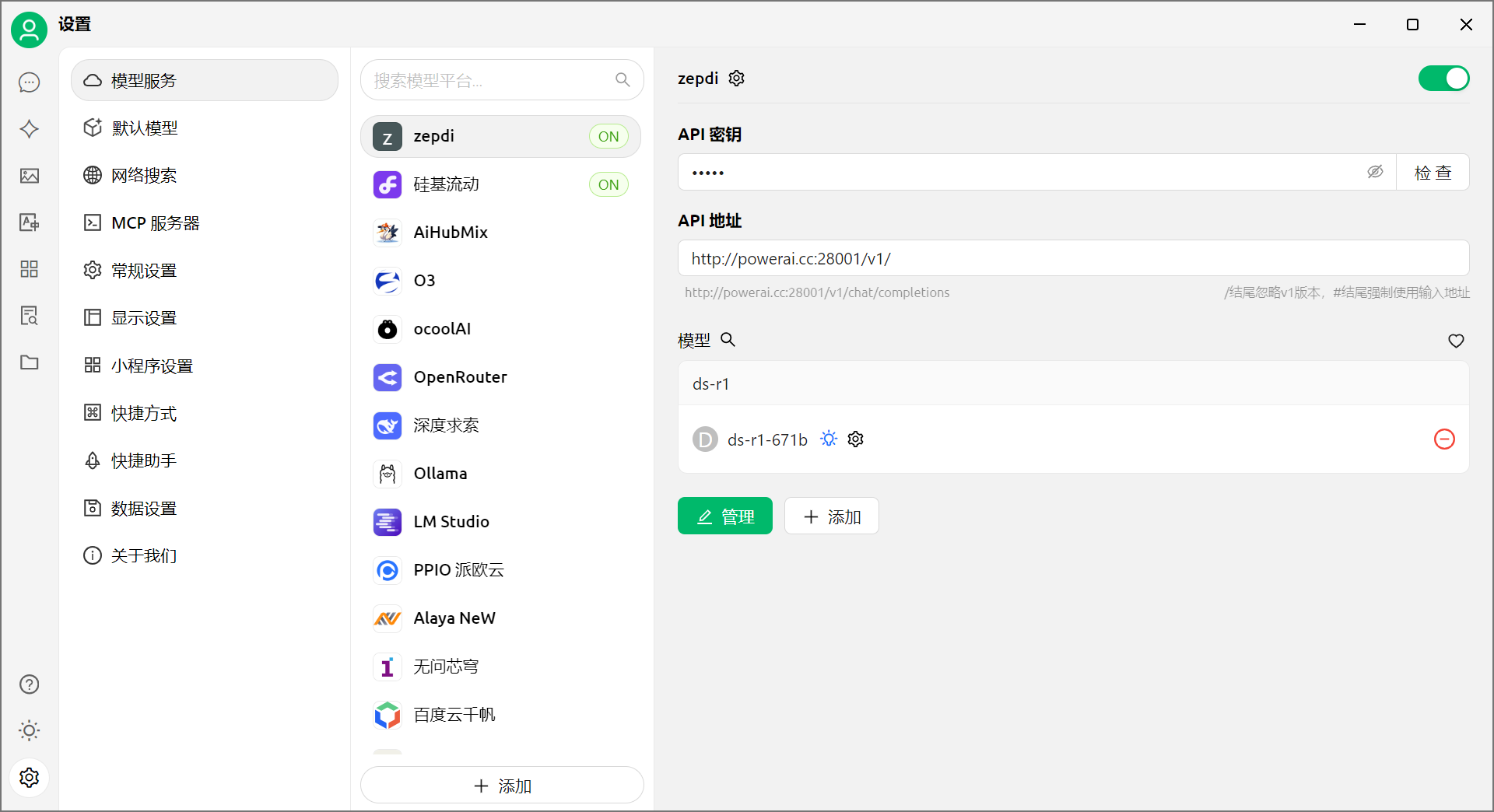
模型ID：ds-r1-671b

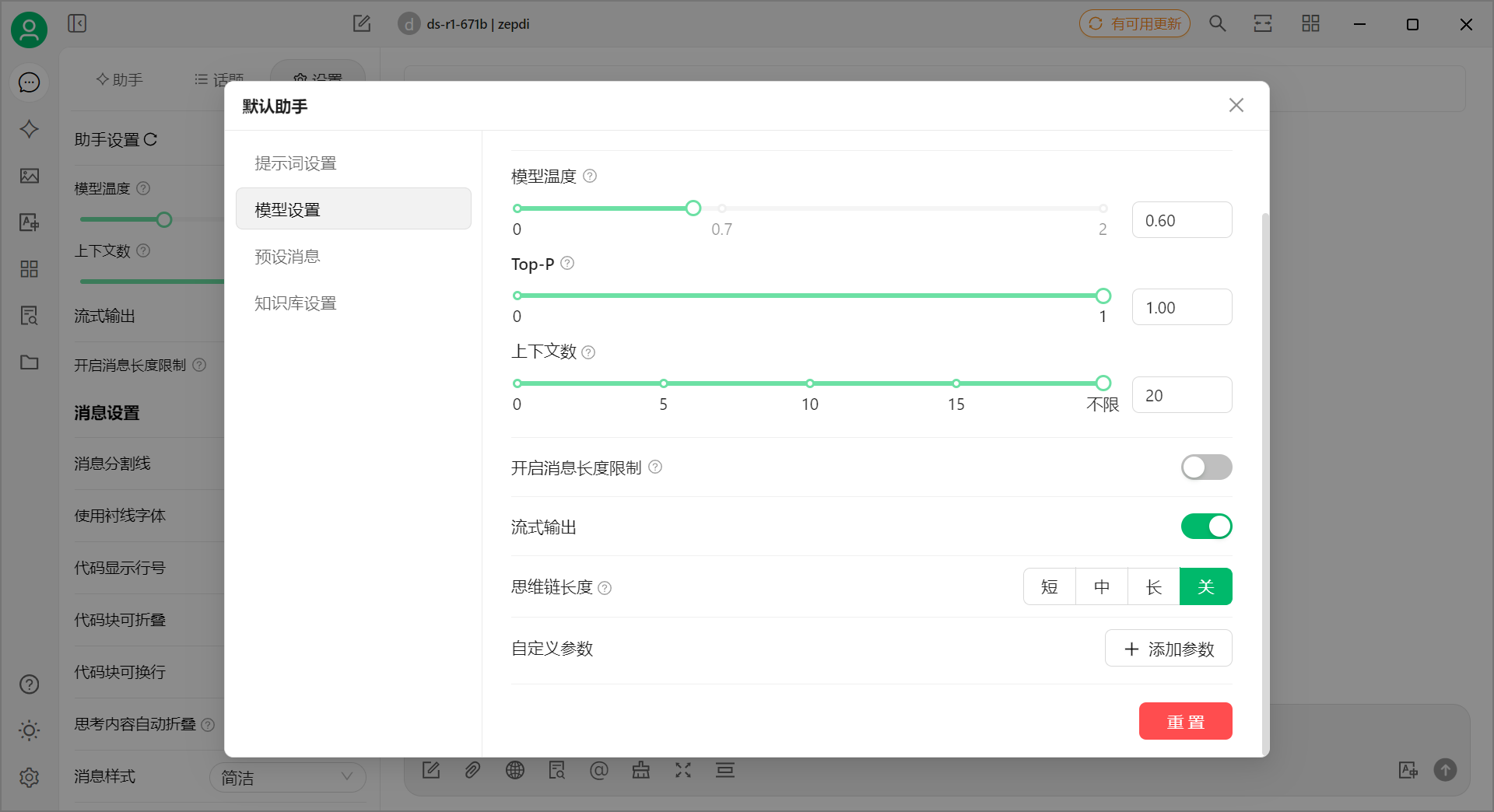
3、对话设置

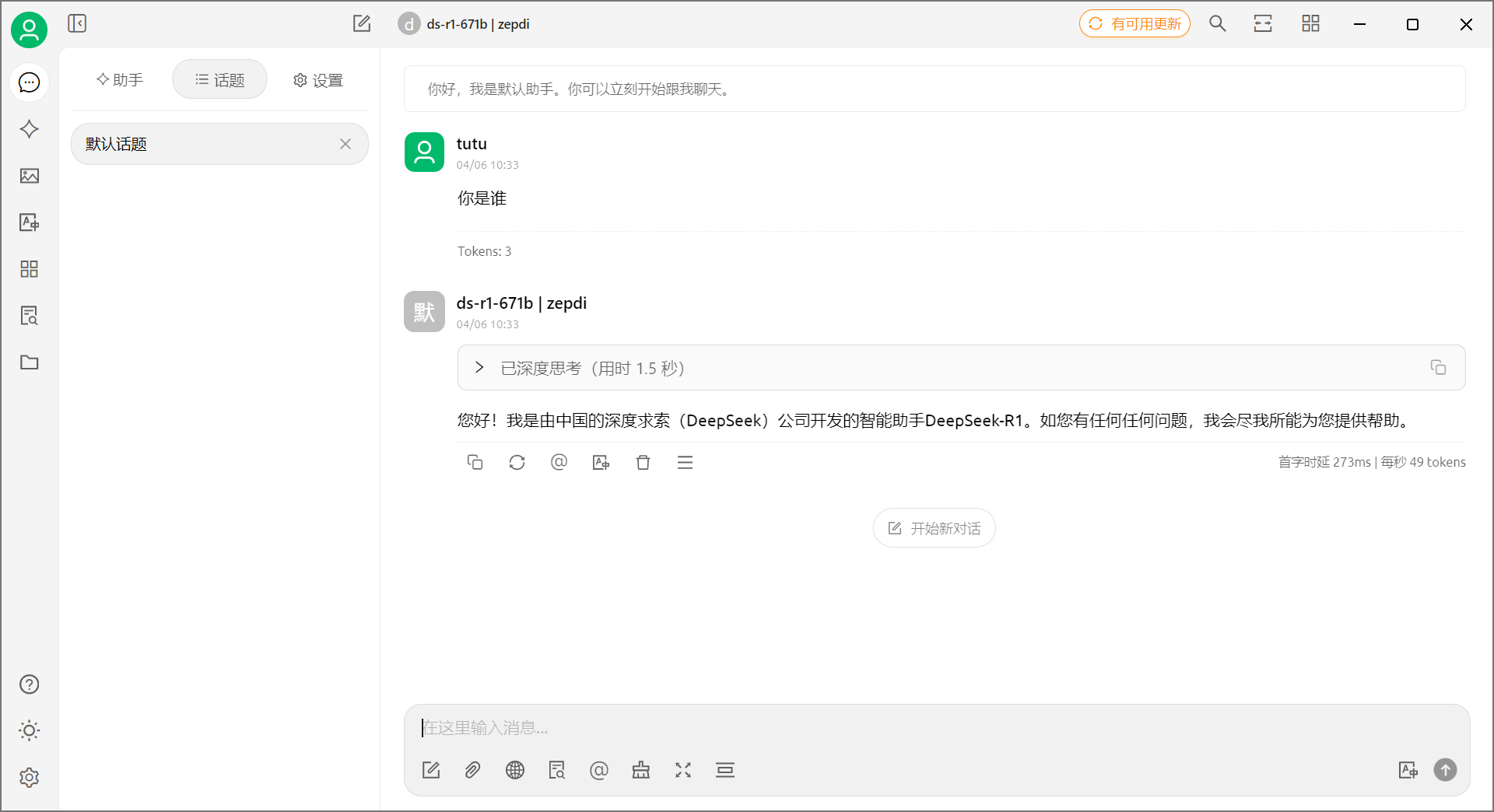
1）左上角，设置->助手设置：

模型温度：0.6

上下文数：不限







注：目前ds-r1模型上下文最大长度65000，最大并发64，开启ep-moe、MTP情况下单线程约30t/s。

(服务器空载功率约2.2kW，单线程推理功率约3.1kW，通过逐步增加负荷方式测得：满负荷推理或训练功率约4.5kW)

