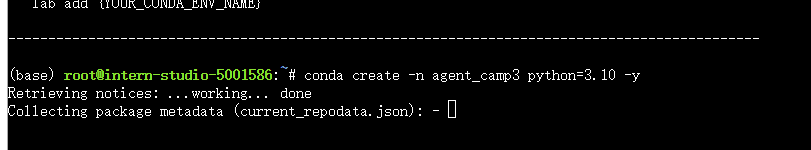
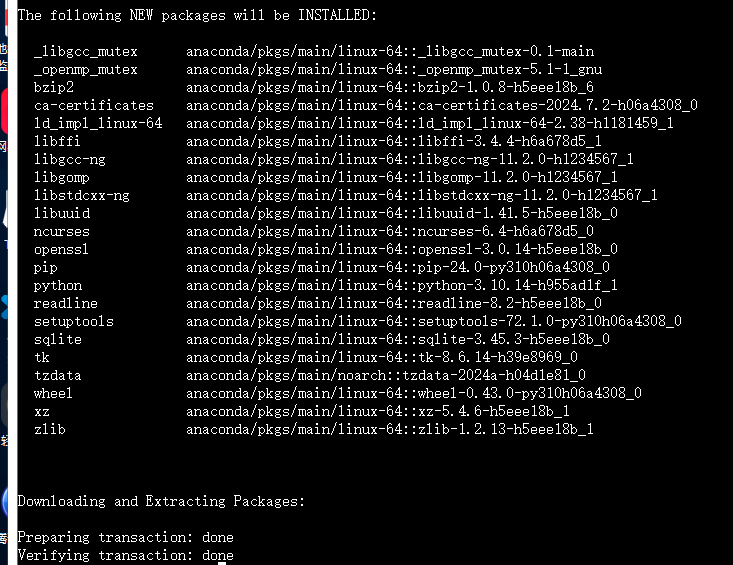
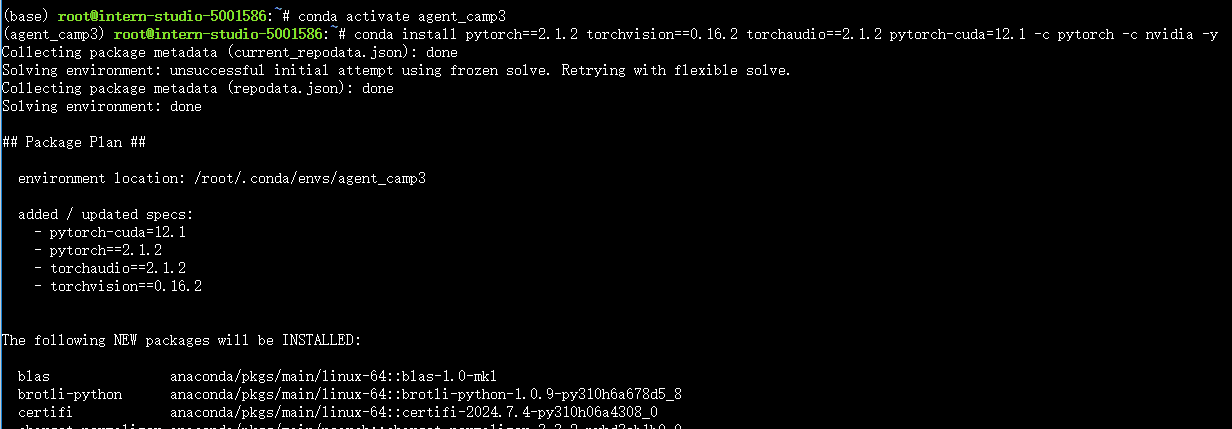
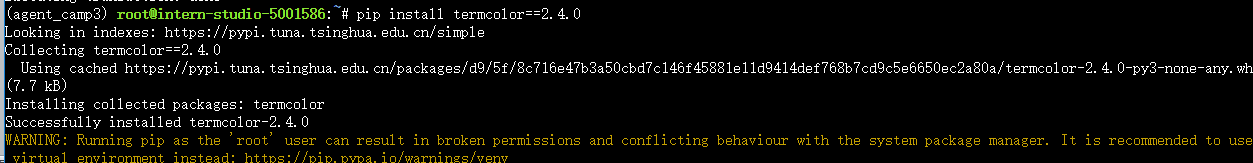
**环境配置**

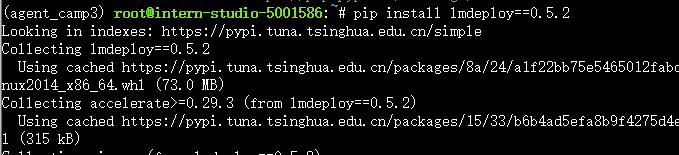


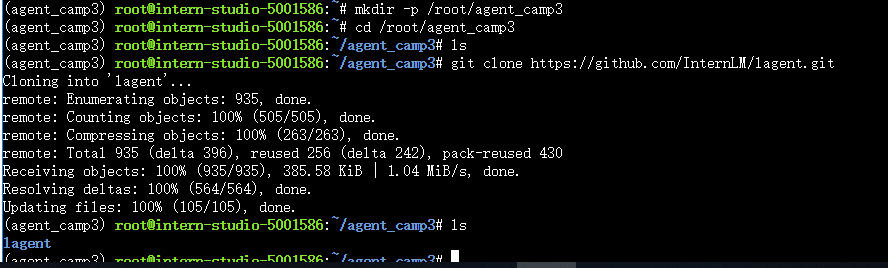


conda install pytorch==2.1.2 torchvision==0.16.2 torchaudio==2.1.2 pytorch-cuda=12.1 -c pytorch -c nvidia –y

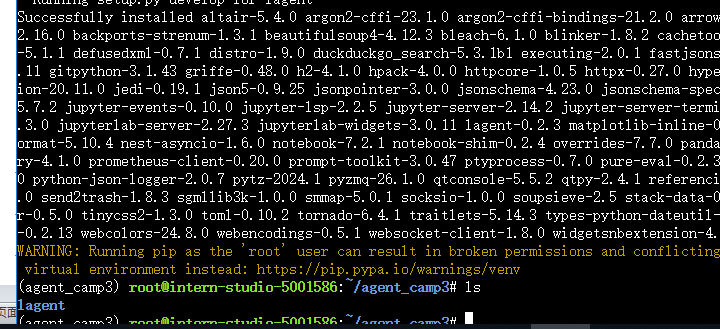




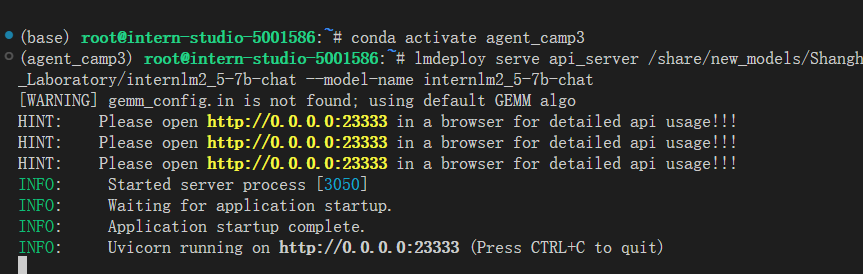


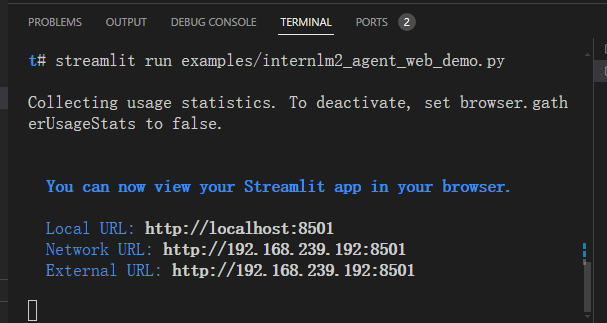




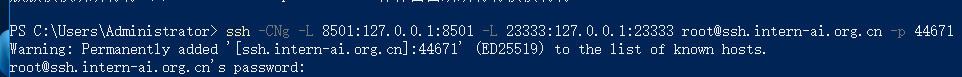


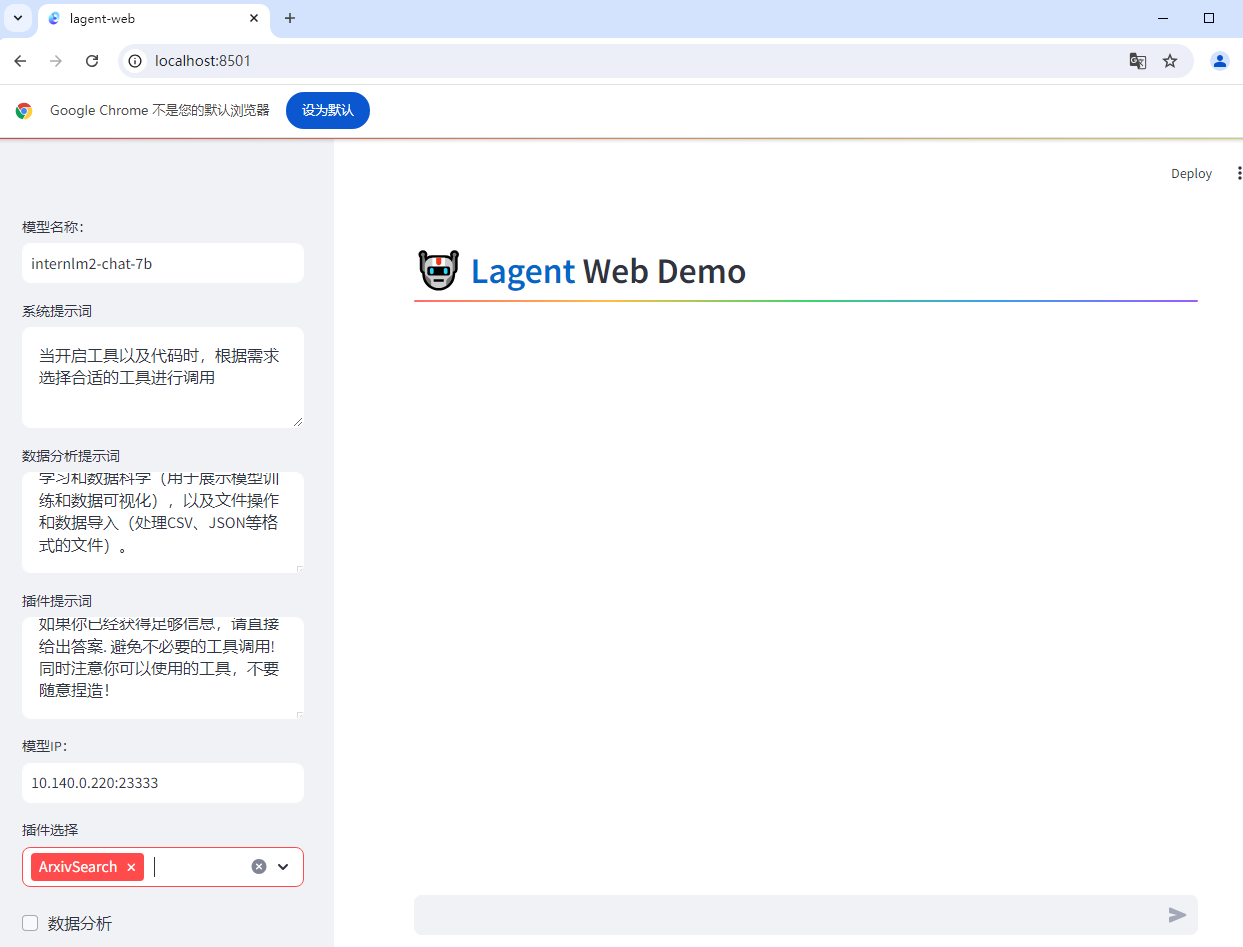
**Lagent Web Demo 使用**





在等待两个 server 都完全启动（如上两图所示）后，在 **本地** 的 PowerShell 中输入如下指令来进行端口映射：









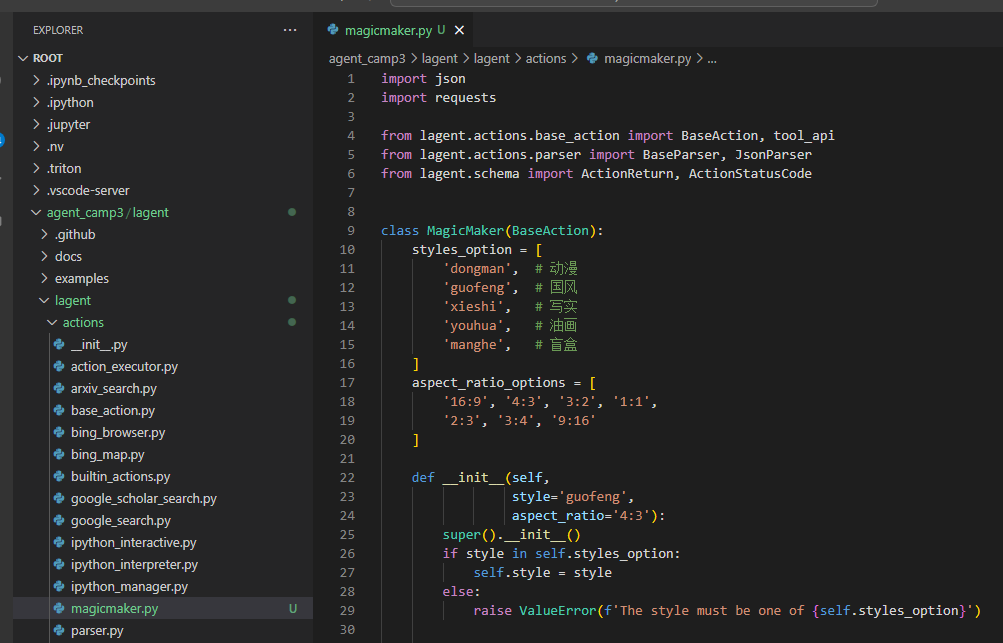
我已经找到了关于 "all you need is attention" 的论文。以下是三篇相关的论文：

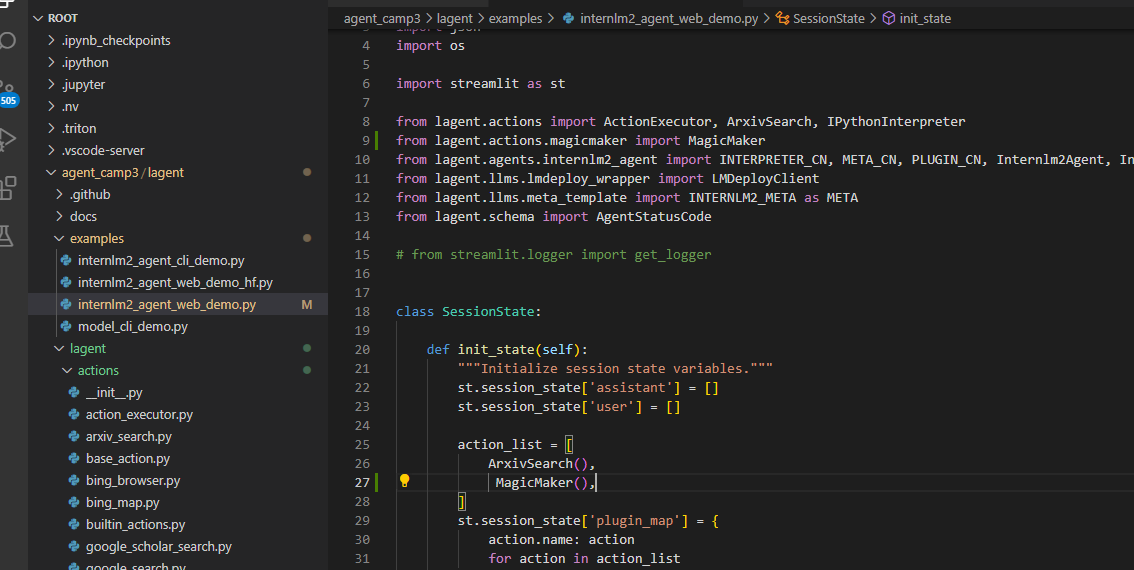
1. 论文标题：Attention Is All You Need But You Don't Need All Of It For Inference of Large Language Models 作者：Georgy Tyukin, Gbetondji J-S Dovonon, Jean Kaddour, Pasquale Minervini 摘要：该论文研究了在大型语言模型（LLM）推理过程中，删除MLP和注意力层对性能的影响。研究发现，删除更深层的注意力层仅略微降低性能，但与删除整个层相比，可以带来更好的速度提升。例如，在13B Llama2模型中，删除33%的注意力层会导致平均性能下降1.8%。
2. 论文标题：All the attention you need: Global-local, spatial-channel attention for image retrieval 作者：Chull Hwan Song, Hye Joo Han, Yannis Avrithis 摘要：该论文提出了一种名为GLAM的全球局部注意力模块，该模块在骨干网络末尾附加，并整合了所有四种形式的注意力：局部和全局、空间和通道。通过空间池化，该模块学习了一种强大的嵌入，用于图像检索。
3. 论文标题：RITA: Group Attention is All You Need for Timeseries Analytics 作者：Jiaming Liang, Lei Cao, Samuel Madden, Zachary Ives, Guoliang Li 摘要：该论文提出了一种名为RITA的时间序列分析工具，该工具使用一种名为组注意力的新型注意力机制来解决时间序列分析中的可扩展性问题。组注意力动态地将对象根据其相似性聚集成一小部分组，并在粗粒度组级别上近似计算注意力。这显著减少了时间和空间复杂度，同时提供了计算注意力的理论保证。

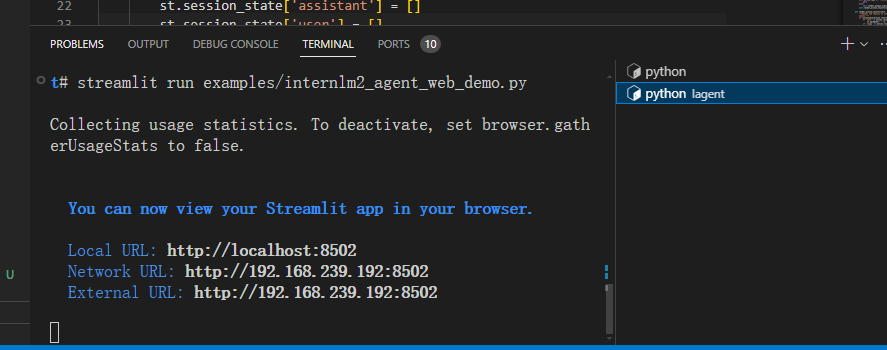
多搜寻了一下，这个应用还真不错！

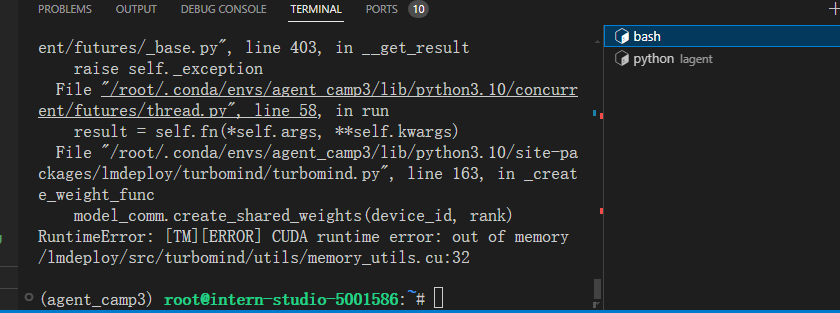
**基于 Lagent 自定义智能体**

下面我们将实现一个调用 MagicMaker API 以完成文生图的功能。

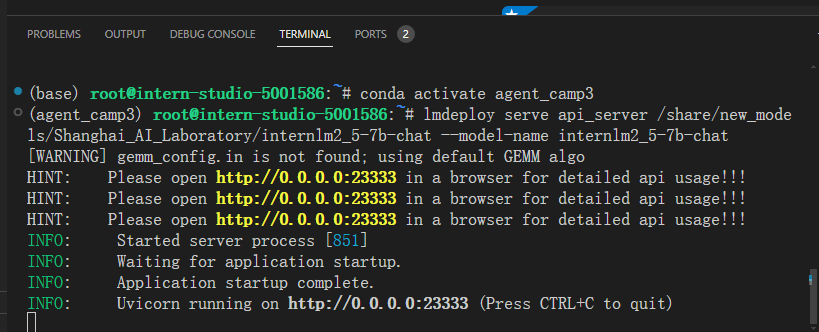




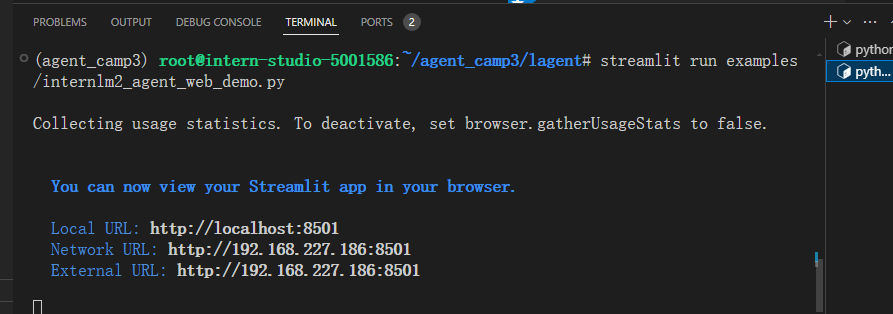




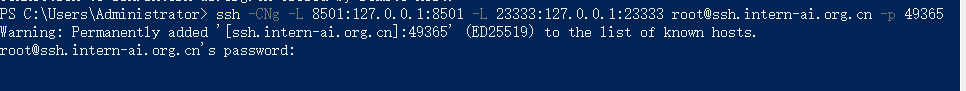
API SERVER没能启动，下来增加GPU使用率。



这次用到100%100的一张A100卡，很快就启动了服务。



重起了开发机，端口号也变了。

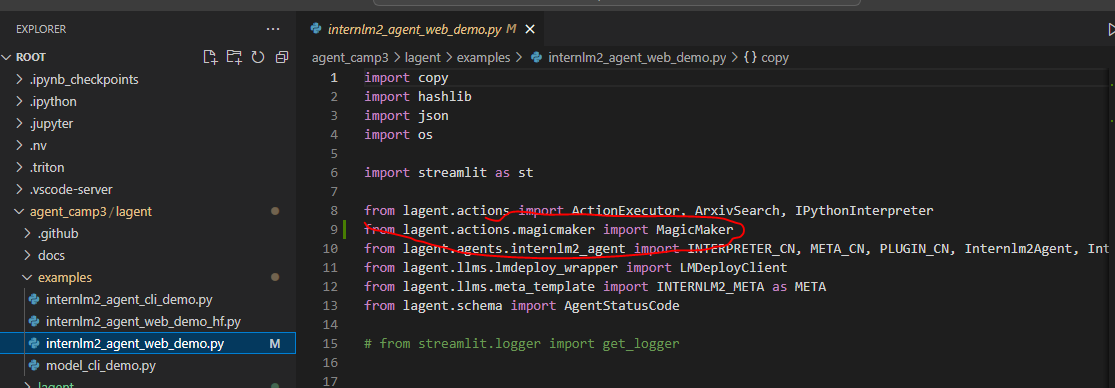


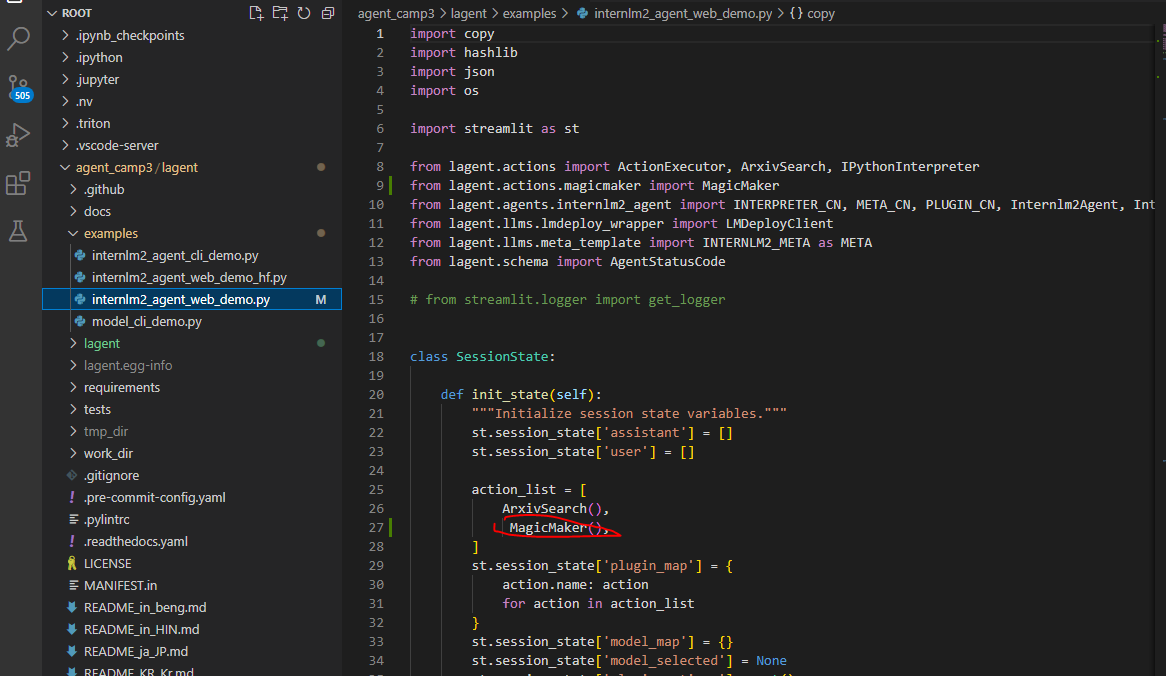


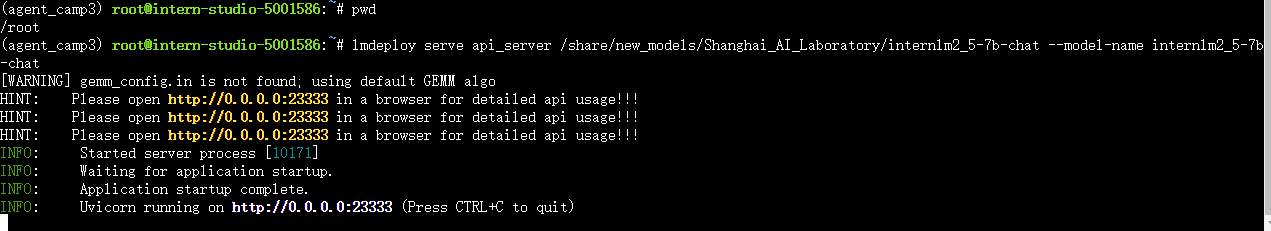


咱这代码有什么问题吧，能部署问题，不给我直接显示生成。重启开发机，

下面重做：用50%的A100







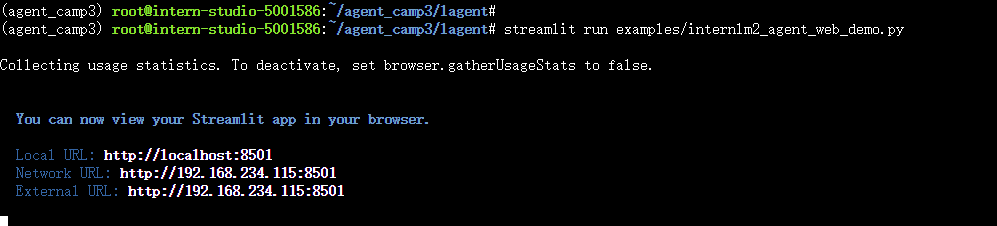
然后，我们在另一个窗口中启动 Lagent 的 Web Demo。

然后，我们在另一个窗口中启动 Lagent 的 Web Demo。

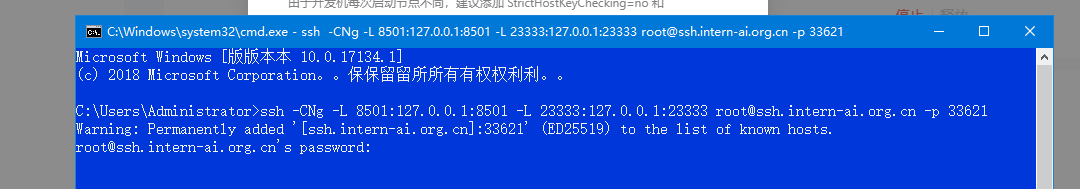
cd /root/agent\_camp3/lagent

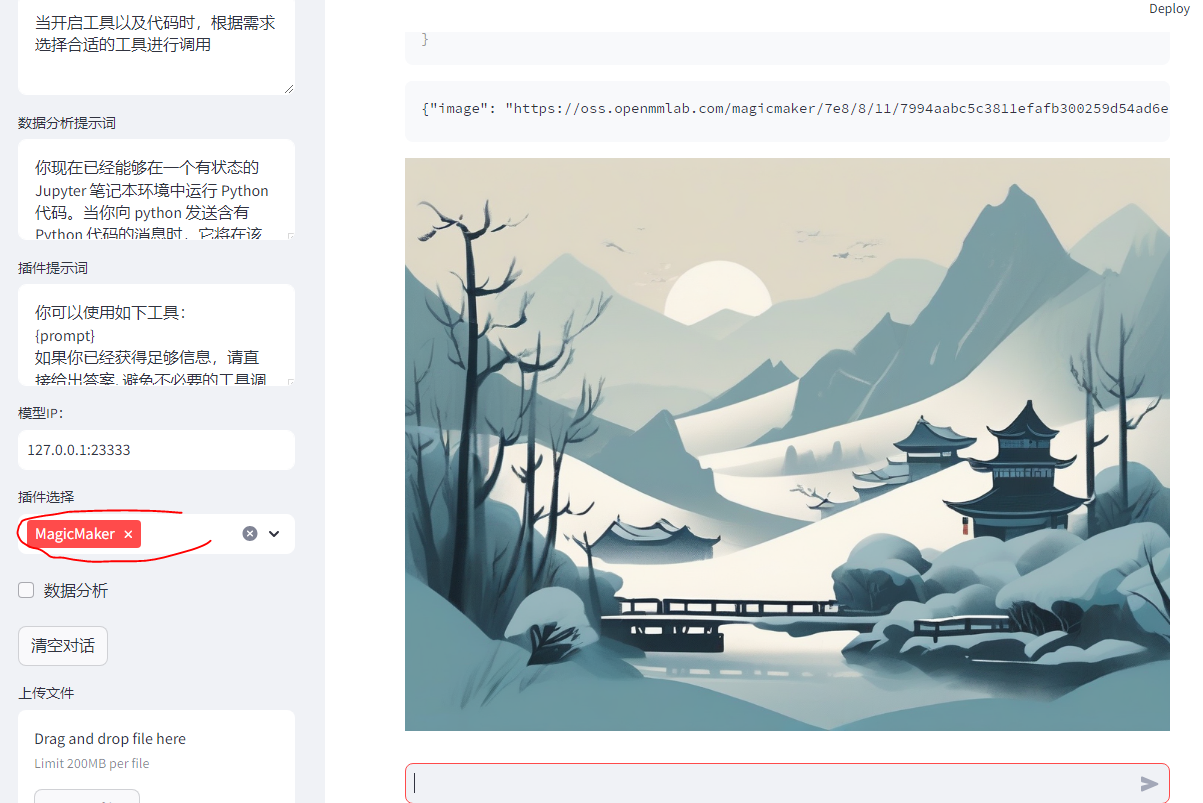
conda activate agent\_camp3

streamlit run examples/internlm2\_agent\_web\_demo.py



本地cmd下，进行端口映射：





终于完成了，上次出错，是因为没有选择插件MagicMaker



