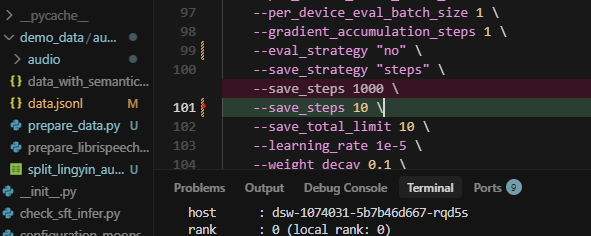
# 模型训练遇到的问题和解决

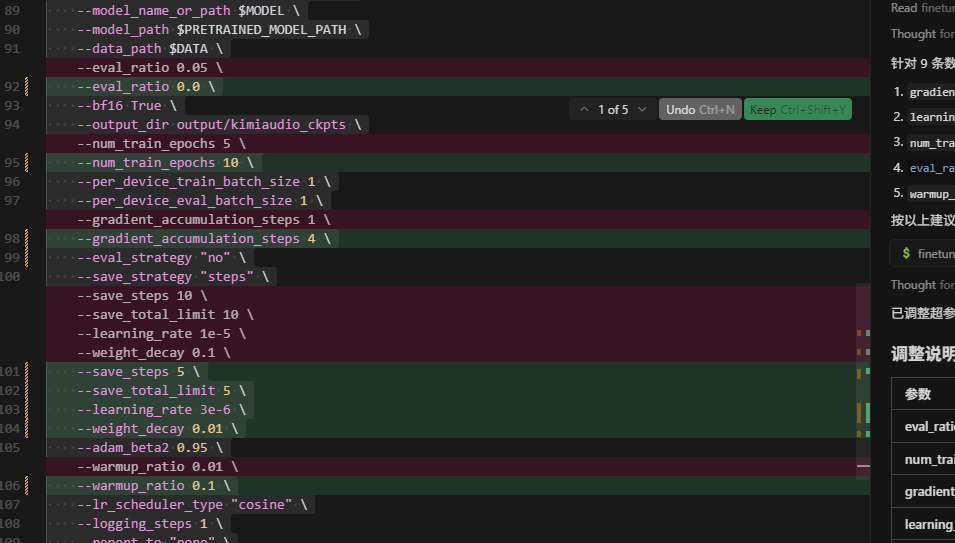
# 1.checkpoint无保存以及参数名更新问题

因为训练集较小只有9条， batch\_size=1，每个 epoch 只有 9 步，5 个 epoch 共约 45 步；而 save\_steps=1000，因此没有触发保存.



模型微调过程中会报错，说有些参数未被识别，查看最新的技术文档发现新版本的 transformers 可能已改为 eval\_strategy

# 2.超参数调整/finetune\_ds.sh



针对 9 条数据的小数据集，建议调整：

gradient\_accumulation\_steps: 1 → 4 或 8（模拟更大 batch，训练更稳定）

learning\_rate: 1e-5 → 3e-6 或 5e-6（降低学习率，降低过拟合风险）

num\_train\_epochs: 5 → 10 或 15（增加训练轮次）

eval\_ratio: 0.05 → 0（数据太少，不分验证集）

warmup\_ratio: 0.01 → 0.1（小数据集使用更多 warmup）

**调整说明**

| **参数** | **原值** | **新值** | **原因** |
| --- | --- | --- | --- |
| **eval\_ratio** | 0.05 | **0.0** | 9条数据分验证集意义不大，  全部用于训练 |
| **num\_train\_epochs** | 5 | **10** | 数据量小，增加轮次 |
| **gradient\_accumulation\_steps** | 1 | **4** | 模拟更大 batch（实际 batch size = 1×4 = 4），  训练更稳定 |
| **learning\_rate** | 1e-5 | **3e-6** | 降低学习率，降低过拟合风险 |
| **weight\_decay** | 0.1 | **0.01** | 降低正则化，适应小数据集 |
| **warmup\_ratio** | 0.01 | **0.1** | 增加 warmup 步数，训练更稳定 |
| **save\_steps** | 10 | **5** | 每 5 步保存（约每半个 epoch 一次） |
| **save\_total\_limit** | 10 | **5** | 只保留最新 5 个 checkpoint，  节省空间 |

**预期效果**

1. 训练更稳定：gradient\_accumulation\_steps=4 等效 batch size 为 4
2. 降低过拟合：更小的学习率和 weight\_decay
3. 更好的学习：更多轮次和更长的 warmup
4. 自动保存：每 5 步保存一次，不会丢失模型

# 3. Nccl警告问题

出现 NCCL 警告，可能卡在分布式通信。数据量很小（9 条），使用 8 个 GPU 进行分布式训练可能导致问题。

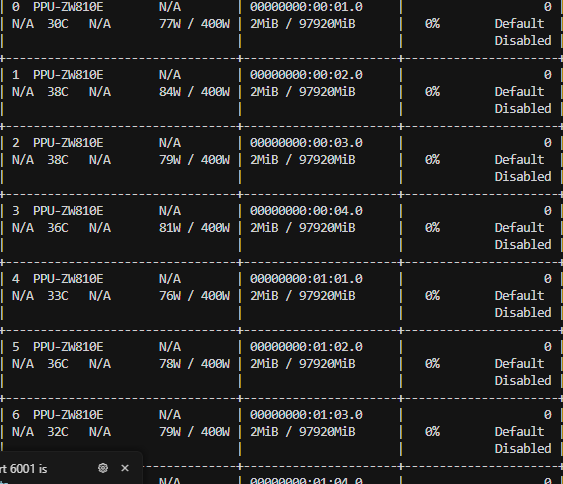
1. 每个 GPU 数据极少（约 1–2 条）
2. 通信开销远大于计算，容易卡在 NCCL 同步
3. DeepSpeed Zero3 在极小的数据上通信效率低

解决方法：



# 4.训练过程中gpu内存过载问题

查看gpu指令： nvidia-smi



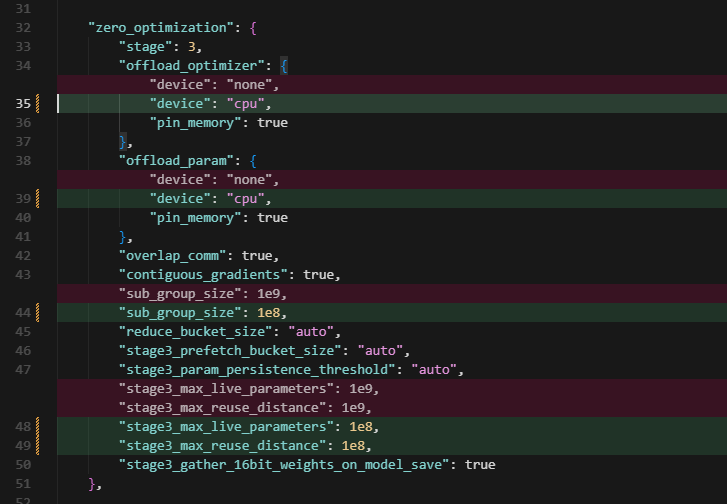
出现 CUDA OOM。问题：

* GPU 总容量：95.62 GiB
* 可用：仅 730.70 MiB
* 已分配：87.03 GiB（PyTorch）+ 7.49 GiB（保留未分配）
* 尝试分配：3.73 GiB 失败

原因是 DeepSpeed Zero3 在单 GPU 上的内存开销过大

## 有两种解决方案：

1. 修改 DeepSpeed 配置：启用 CPU offload、减小 model\_max\_length在训练脚本中减小序列长度、减小 batch size



1启用 CPU offload：将 offload\_optimizer 和 offload\_param 的 device 从 "none" 改为 "cpu"

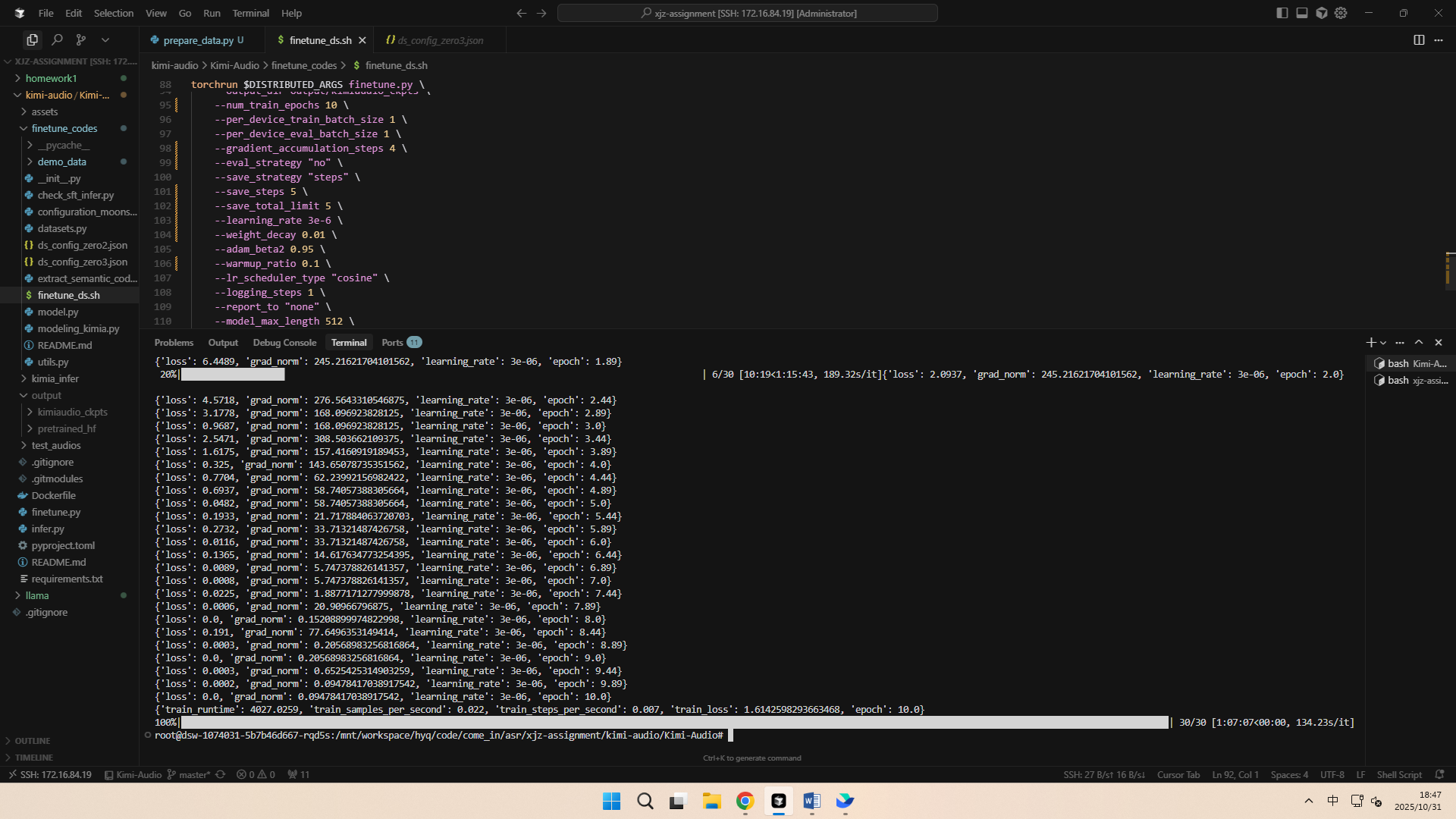
2减少内存相关参数：将 stage3\_max\_live\_parameters、stage3\_max\_reuse\_distance 和 sub\_group\_size 从 1e9 减小到 1e8

**这样会将优化器状态和部分模型参数 offload 到 CPU，显著降低 GPU 显存占用。**

1. 分布式训练：由单gpu转为多gpu（但对于小样本训练，多gpu训练又会导致通信问题）

# 5.结果

## 模型训练结果



## 模型推理结果

