# 模型迁移第二阶段提交文档

## 队伍名称：aabb

## 队伍成员：苏仪

## 使用的优化方法：

1、ops改为mint。

2、改变rope中的一些操作，如取值，乘法等。

3、加了rms\_norm大算子和flash\_attention大算子。

4、改变了处理attention mask和position ids的逻辑。

5、改变了prepare input ids的逻辑。

6、rotary embedding以及其他算子中的取固定长度的切片操作改为mint.narrow。

7、Moe部分全部重写，舍去了所有for循环，切片，index加和nonzero等操作。

## 结果展示

### 环境

Ascend-snt-9b2，24核，192G内存，64G显存。

在华为云贵阳一mindspore2.3.0的镜像进去之后:

git clone https://github.com/mindspore-lab/mindnlp/

git checkout bb4876ec53f20f8224960c6bdee17f0fe80db412

pip install -r requirements/requirements.txt

pip uninstall soundfile

pip install mindspore==2.3.1

### Mixtral结果展示

命令：

cd mindnlp

unset RANK\_TABLE\_FILE

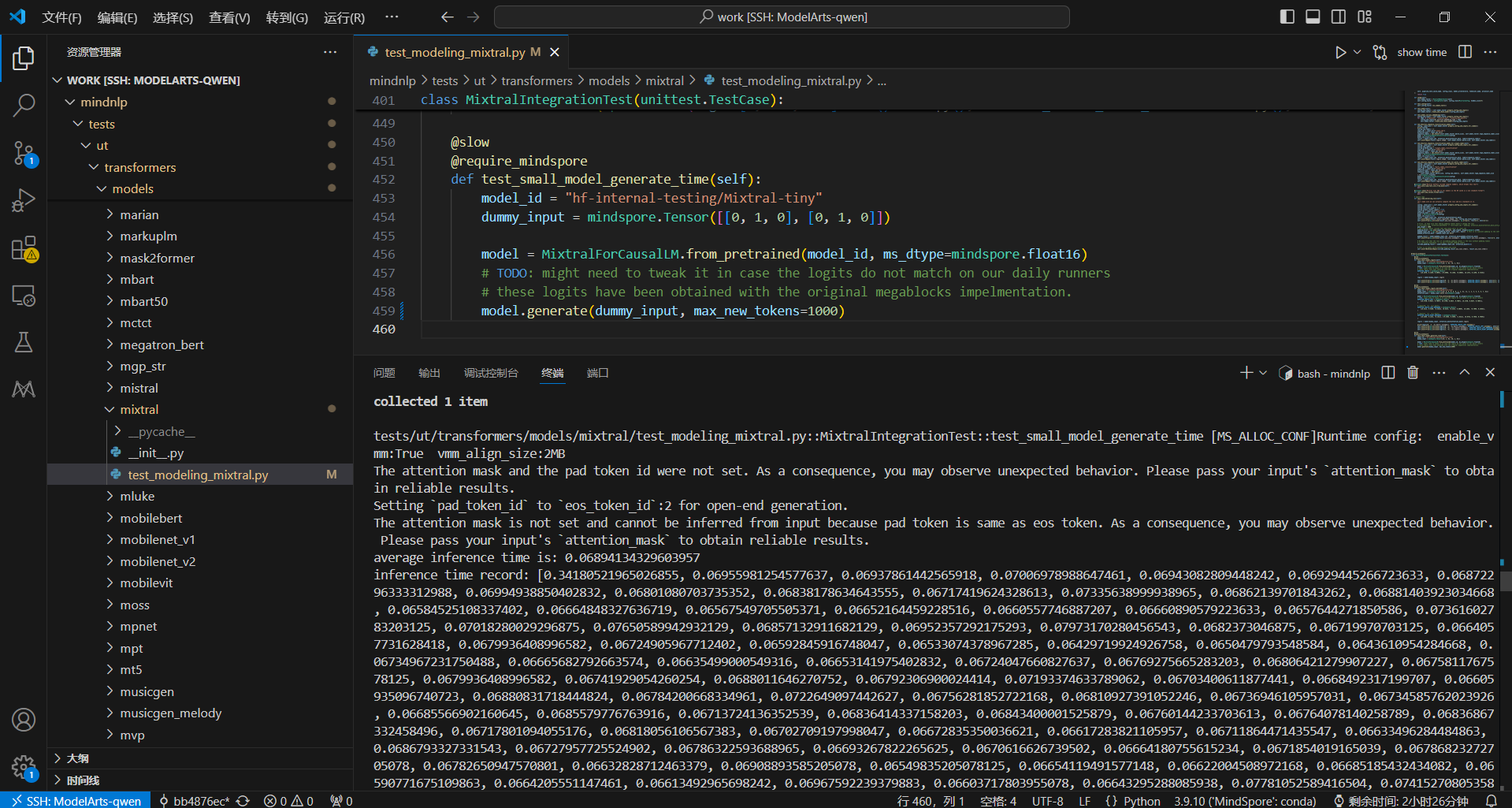
export INFERENCE\_TIME\_RECORD=1

export RUN\_SLOW=1

taskset -c 0-23 pytest -vs tests/ut/transformers/models/mixtral/test\_modeling\_mixtral.py::MixtralIntegrationTest::test\_small\_model\_generate\_time

#### 原始速度评测（ut中max\_new\_tokens改为1000）

#### average inference time is: 0.06894134329603957



#### 修改后速度评测（ut中max\_new\_tokens改为1000）

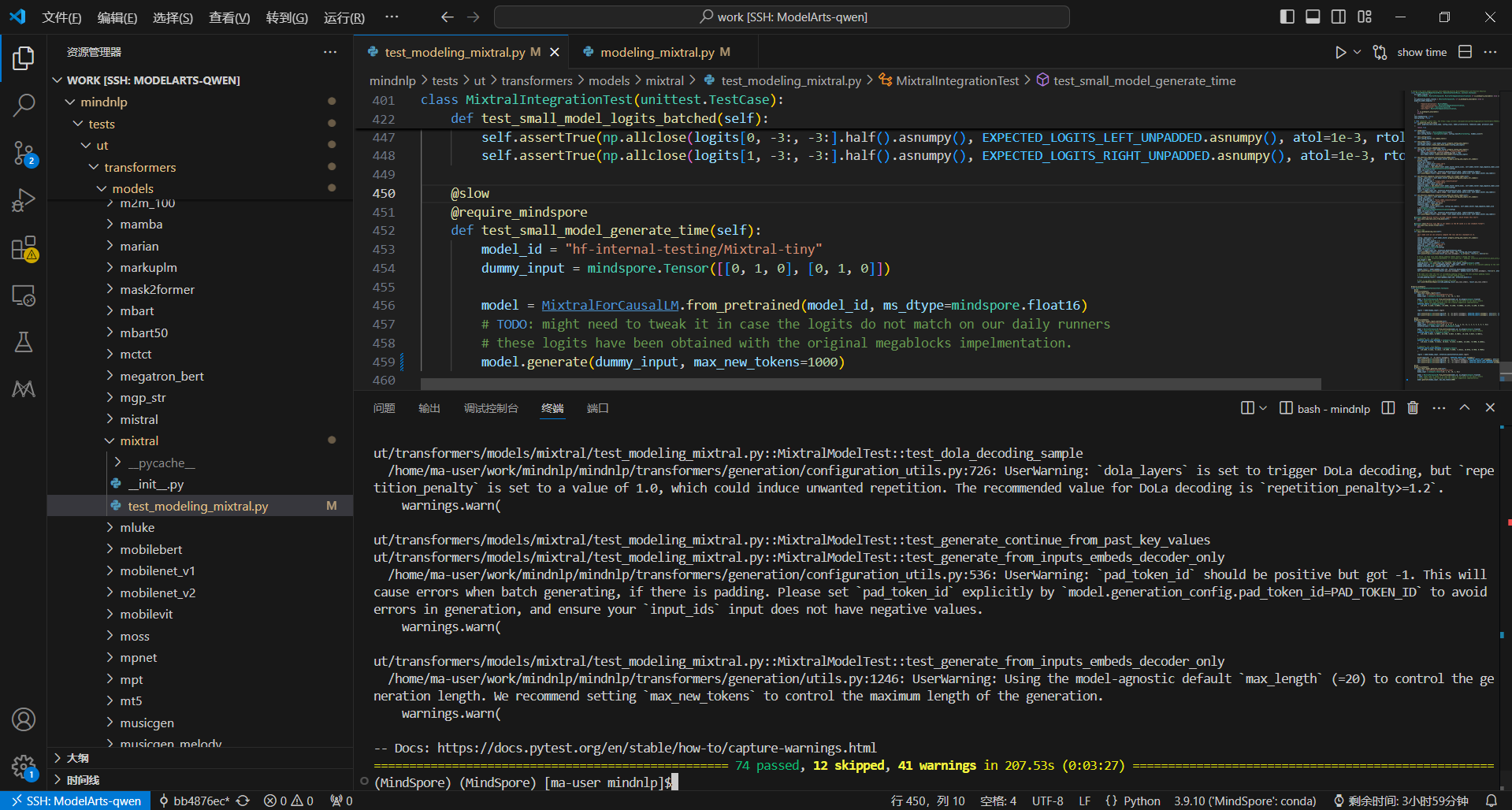
#### average inference time is: 0.01021228514395438 (提升6.75倍)

#### 

#### 修改后精度验证

export INFERENCE\_TIME\_RECORD=0

pytest -vs tests/ut/transformers/models/mixtral/test\_modeling\_mixtral.py



### Qwen结果展示

命令：

cd mindnlp

unset RANK\_TABLE\_FILE

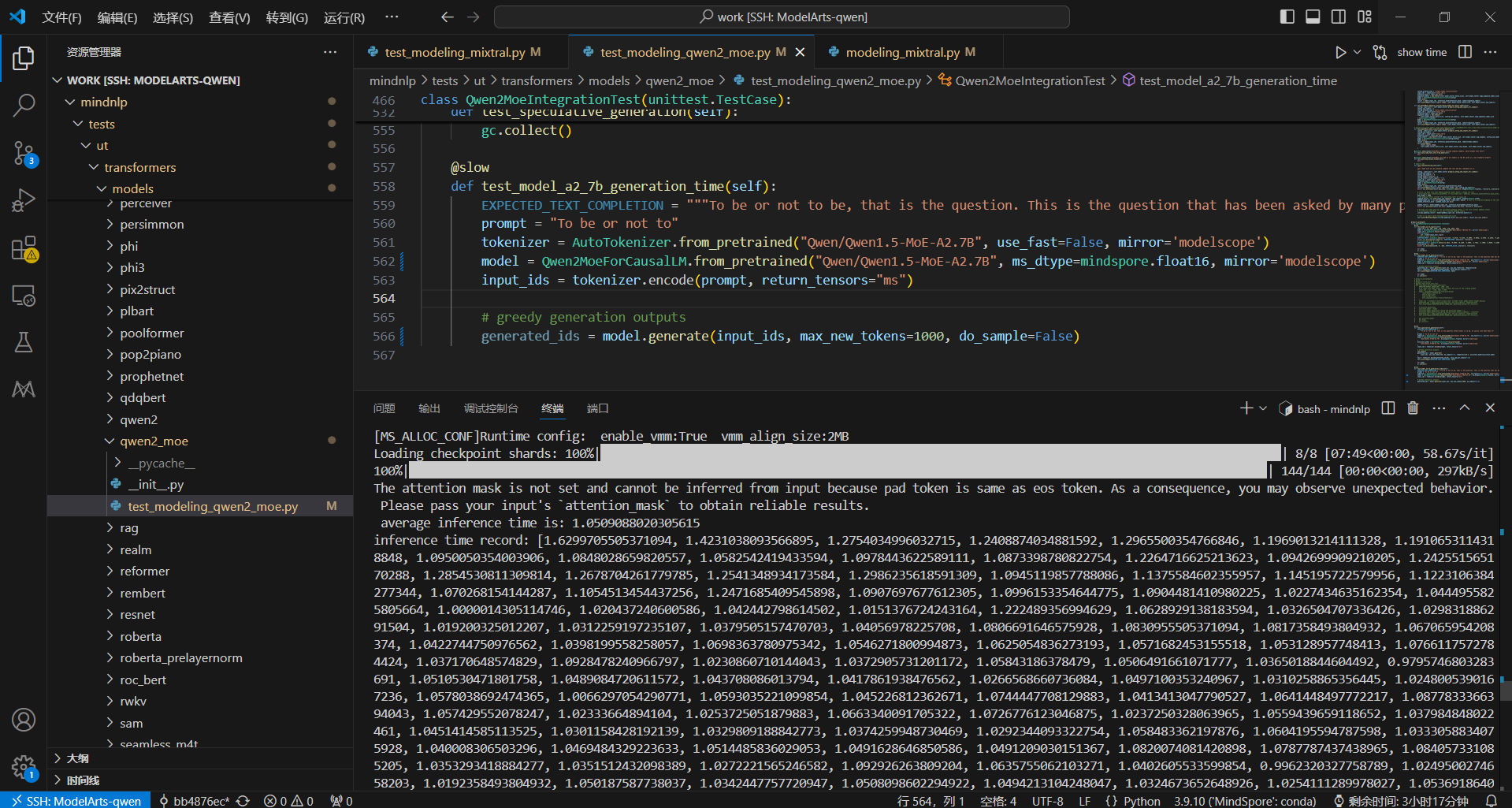
export INFERENCE\_TIME\_RECORD=1

export RUN\_SLOW=1

taskset -c 0-23 pytest -vs tests/ut/transformers/models/qwen2\_moe/test\_modeling\_qwen2\_moe.py::Qwen2MoeIntegrationTest::test\_model\_a2\_7b\_generation\_time

#### 原始速度评测（ut中max\_new\_tokens改为1000， 且fp16）

#### average inference time is: 1.0509088020305615



#### 修改后速度评测（ut中max\_new\_tokens改为1000）

#### average inference time is: 0.08974514900146423 (提升11.7倍)

#### 

#### 修改后精度验证（经群里说明，跳过投机解码）

export INFERENCE\_TIME\_RECORD=0

pytest -vs tests/ut/transformers/models/qwen2\_moe/test\_modeling\_qwen2\_moe.py

