FAQ：

启动服务过程：

1. 如何启动serving服务？

* 修改serving\_config.py配置文件，参考《MindSpore\_Serving\_2.1分支接口文档》第一章节-模型及组batch配置
* 执行python start.py一键启动agents和serving，对应生成agent.log和server.log

1. serving启动过程报错 Connection refused

* 检查端口号是否被占用
* 检查server.log中的agent ports和agent.log中的agent ports是否一致，不一致的话，排查PYTHONPATH是否有问题

1. agent启动中报错’last process is running’

* 检查卡是否被占用

1. agent启动中报错rank\_table不匹配

* 检查serving\_config.py的’device’字段是否和ctx\_path/inc\_path配置中的rank\_table相匹配

1. agent启动中报错’predict failed’或者’input size is wrong’

* 检查agent.log中打印出来的input shape/type是否和提供模型的入参一致

1. agent启动中报错’build\_from\_file failed’

* 检查ctx\_path/inc\_path配置的动态设置是否和’dyn\_batch\_size’以及’zactivate\_len’字段相对应

1. serving启动中报错”b’failed’ there exists another connected serving now”

上一次serving异常退出（没有使用ctrl+c 或者 kill -15），需要重新拉起agent

请求服务过程：

1. 发送请求无响应

* 检查发送请求的接口以及ip:port是否正确
* 检查server.log 是否收到了发送的请求，是的话检查output/agent\_0.log（logging设置为info级别）是否有推理的日志。如果agent侧没有开始推理，检查output/server\_app.log中是否有serving相关的处理（logging设置为info级别），如果serving也没有任何处理，请联系开发人员；如果agent侧和server\_app.log都在运行，那么可能是调用非流式推理时设置了较大的max\_new\_token值，等待整体的返回时间偏长，可通过如下方式确认系统正常运行：1. 查看serving日志，请求接受则正常；2. 查看agent的日志，查看单token的推断时间，如果持续打印npu推断时间，且时长合理（50-60ms）则正常；此时可以考虑使用流式返回接口

1. agent正常启动，发起请求后agent日志报错：agent/agent\_multi\_post\_method.py \_post\_sampling\_topk\_npu IndexError: list index out of range：

后处理模型有误，解决：

* 需要重新导出后处理模型，方法：工程目录下运行post\_sampling\_model.py，生成后处理模型，；
* 检查config/serving\_config.py，argmax\_model，topk\_model设置为1.生成模型的路径，注意argmax、topk要对应；