

# 智能审批模型国产化推理适配工作汇报

汇报人：智能审批研发团队

汇报对象：信贷管理部领导

汇报时间：约8分钟

## 开场白

各位领导好：

今天由我代表智能审批研发团队，向大家汇报一项重要的基础设施保障工作——核心AI模型的国产化推理适配。

简单来说，就是把原来跑在进口芯片上的智能审批模型，成功迁移到了国产芯片上，而且性能、精度都达到了生产标准。这项工作直接关系到审批系统的供应链安全和长期稳定运行。

下面我将按照项目背景、实施方案、攻坚过程和最终成果四个部分，向大家作简要汇报。

## 第一部分：为什么要做这件事？

### 1.1 项目背景

目前，咱们生产环境所有的AI推理服务都部署在英伟达A100芯片上，使用Triton推理引擎进行调度。这套架构虽然成熟，但存在两个不可忽视的风险：

**第一，供应链风险。** 进口高端GPU受国际贸易政策影响，存在断供和限售的可能。一旦硬件无法持续采购，后续的扩容、替换都会受到制约。

**第二，合规压力。** 随着国家对关键信息基础设施自主可控的要求越来越高，金融核心系统使用国产化算力平台是大势所趋，也是监管鼓励的方向。

因此，我们在今年一季度启动了专项适配工作，目标很明确：**在不改变模型效果的前提下，把核心审批模型从进口平台迁移到国产平台。**

### 1.2 项目目标

这次工作聚焦在“推理侧”，也就是模型已经训练好了，只把它搬到新硬件上跑起来。核心目标有三个：

1. **模型能跑**：把审批系统里最核心的BERT-Small模型，完整迁移到国产昇腾910B3平台上，覆盖文本分类和命名实体识别（NER）两类任务——这两类任务支撑着咱们智能审批里的材料解析和字段抽取。
2. **效果不降**：迁移后的模型输出，与原平台相比，准确率差异控制在千分之五以内，输出相似度达到99.9%以上。说直白点，客户感受不到任何变化。
3. **速度够快**：新平台的推理速度要达到原平台的80%以上，响应延迟增幅不超过20%。确保审批体验不会变慢。

## 第二部分：我们选了什么方案？

### 2.1 硬件选型：昇腾910B3

经过调研，我们选定了华为昇腾910B3作为目标平台。给大家报几个关键数字，方便理解：

- **算力**：半精度浮点算力313万亿次/秒，与A100基本持平；整型算力626万亿次/秒，甚至略高。
- **显存**：64GB高速显存，对于我们当前使用的BERT-Small模型完全够用。
- **功耗**：400瓦，与A100相同，不需要改造机房供电。

一句话总结：**这块国产芯片在算力上不输A100，完全能撑起我们的审批模型。**

### 2.2 软件栈：LitServe + ACL

硬件定了，软件怎么搭？原来的Triton推理引擎是英伟达生态的产物，在国产芯片上无法直接使用。

我们最终选择了**LitServe + ACL**的组合：

- **ACL**：昇腾官方的底层运行时，相当于国产芯片的"驱动程序"，负责加载模型、调度计算。
- **LitServe**：基于Python的高性能推理服务框架，相当于国产化的"Triton替代品"，负责接收请求、批量处理、对外暴露API。

这个方案的优势是**轻量、可控、部署快**。团队预估3个工作日就能完成核心开发，实际也验证了这个判断。

## 第三部分：我们是怎么做的？

### 3.1 总体路线

整个迁移过程可以概括为"五步走"：

1. **导出**：把PyTorch模型导出为通用的ONNX格式，相当于把模型"翻译成世界语"。
2. **编译**：用昇腾的ATC编译器把ONNX转成昇腾专用的OM格式，同时开启算子融合和精度优化。
3. **封装**：用ACL接口编写推理逻辑，包括数据搬运、计算执行、结果回传。
4. **服务化**：用LitServe把推理逻辑包装成Web服务，支持自动批处理和并发扩展。
5. **验证**：对比新旧平台的输出结果，确保精度和性能达标。

## 3.2 实施过程

整个专项工作由2名工程师并行推进，实际耗时3个工作日，与预估的5人天基本吻合。具体工作包括：

- **环境搭建**：完成驱动安装、工具链部署、容器化配置。
- **模型转换**：完成BERT-Small的ONNX导出和OM编译。
- **服务开发**：完成ACL推理封装和LitServe服务化改造。
- **精度对齐**：通过逐层对比激活值，定位并修复了LayerNorm、Gelu等算子的精度差异。
- **压测调优**：调整批处理大小和内存策略，最终性能达到预期。

## 第四部分：遇到了哪些困难？怎么解决的？

技术迁移不可能一帆风顺，我们主要攻克了三类问题：

### 4.1 算子兼容性问题

不同芯片支持的"指令集"不完全一样。比如原模型里用到的 `index_add_` 操作，昇腾芯片不支持"原地修改"版本，我们把它改成了"非原地修改+复制"的等效写法；再比如LayerNorm的方差计算，两家厂商的实现细节有差异，我们通过统一导出前的实现方式，消除了偏差。

### 4.2 精度对齐问题

FP16混合精度计算在不同硬件上的舍入行为存在微小差异，深层网络会累积放大。我们采用了"逐层对比"的方法，从输入层开始一层一层排查，最终把端到端精度差异控制在了0.1%以内。

### 4.3 性能调优问题

昇腾910B3的显存带宽比A100低约21%，这意味着"搬运数据"相对慢一些。我们通过增大批处理大小、预分配内存池、开启算子融合等手段，把单次推理延迟控制在A100的94%，吞吐量达到了92%。

# 第五部分：最终成果如何？

## 5.1 精度验证

我们用相同的审批样本，同时在国产平台和原平台上跑了一遍：

- 输出相似度：99.97%，远超99.9%的目标。
- 最大误差：0.00021，在可接受范围内。

下游任务表现：

- 文本分类：准确率94.32%，与原平台相差仅0.06个百分点。
- 命名实体识别：准确率96.15%，与原平台相差仅0.08个百分点。

结论：模型效果没有任何实质性下降，审批质量完全有保障。

## 5.2 性能验证

在单卡、并发100请求的压力测试下：

- 吞吐量 (QPS)：国产平台达到1659次/秒，是A100基线的92%。
- P99延迟：11.96毫秒，是A100基线的106%，增幅仅6%，远低于20%的红线。
- 显存占用：0.56GB，比A100还低25%。

结论：审批响应速度没有明显变慢，用户体验不受影响。

## 5.3 总体评价

维度	目标	实际	状态
模型迁移	完成BERT-Small	已完成	✅ 达标
精度差异	< 0.5%	0.08%	✅ 达标
推理延迟	< 120%	106%	✅ 达标
吞吐量	> 80%	92%	✅ 达标

**BERT-Small模型已完成上线评审，具备投产条件。**

## 第六部分：后续计划

这次适配不仅解决了一个模型的迁移问题，更重要的是建立了一套可复用的国产化推理能力。接下来我们重点推进四件事：

- INT8量化上线**：进一步降低延迟和显存占用，预计吞吐量还能再提升76%。
- 模板化推广**：把这次的经验沉淀为标准模板，后续新模型适配周期可从5天缩短到2天。
- 多卡并行验证**：测试8卡环境下的负载均衡，提升整机吞吐，为业务高峰做准备。
- 监控体系完善**：补齐Prometheus监控和Grafana大盘，达到与原有平台同等的运维可视性。

## 结语

各位领导，这次国产化适配工作的顺利完成，意味着咱们的智能审批系统在供应链安全和自主可控方面迈出了关键一步。

我们证明了：**国产芯片完全有能力支撑金融核心AI模型的生产推理，而且在效果和性能上都能达到进口平台的同等水平。**

团队已经做好了持续推广和深度优化的准备，确保国产化替代工作平稳、高效、无感地推进。

以上汇报，请各位领导审阅。谢谢！

### 附录：演讲时间分配建议

章节	建议时长	重点
开场白	30秒	点明主题和价值
第一部分：为什么要做	1分30秒	风险、合规、目标
第二部分：选了什么方案	1分钟	硬件算力、软件选型
第三部分：怎么做的	1分30秒	五步路线、实施过程
第四部分：困难与解决	1分30秒	三类问题、解决思路
第五部分：最终成果	2分钟	精度、性能、达标结论
第六部分：后续计划	1分钟	四件事
结语	30秒	总结升华

章节	建议时长	重点
总计	约8分钟	—