惠每科技 | Dr. Mayson
Clinical Decision Support System

基于Sapphire Rapids和AMX实现医疗大模型推理加速

CONTENTS
目录

公司介绍

医疗人工智能解决方案提供商——北京惠每云科技有限公司（简称“惠每科技”）成立于2015 年。秉承“质量即生命”理念，惠每科技组建了高年资医生、医学博士、计算机博士、资深算法专家为核心的研发团队，致力于利用 CDSS 技术提升医疗质量，守卫患者安全。

惠每医疗人工智能解决方案在医院端的核心应用 Dr.Mayson，融入了 PDCA 过程管理和 CDSS 技术，通过实时数据分析与事中智能提示，在临床诊疗决策、病案首页与运行病历、单病种质量管理控制与数据上报、临床诊疗风险预警、DRG/DIP 费用管理等环节形成质控闭环，有效提升医疗质量。产品线覆盖肿瘤、呼吸、心脑血管、ICU 等专科领域，和 58 个国家重点监测单病种。

截至 2023年10月，惠每医疗人工智能解决方案落地 近700 家医院，包括43家复旦版TOP100 医院，助力近 170 家医院通过“电子病历”“互联互通”高级别评审，为医疗质量与患者安全构筑智能防护盾牌。

应用案例

应用案例

- 以病案病历质控为例-

医生端 | 病案室端 | 管理端

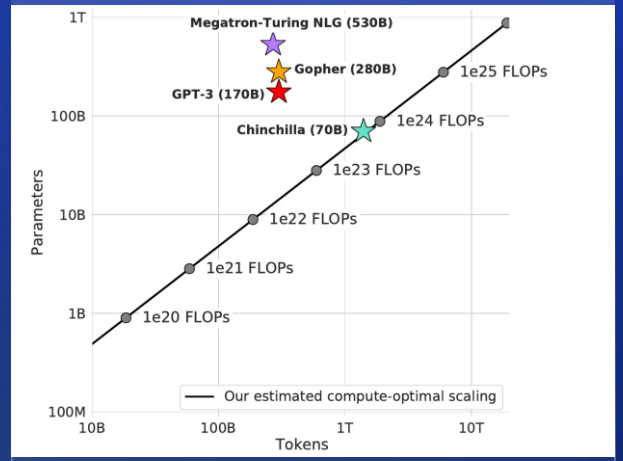
- 利用AI自动执行院内三级质控模式，实现管理关口前移和及时干预

- 运行病历质控关口前移

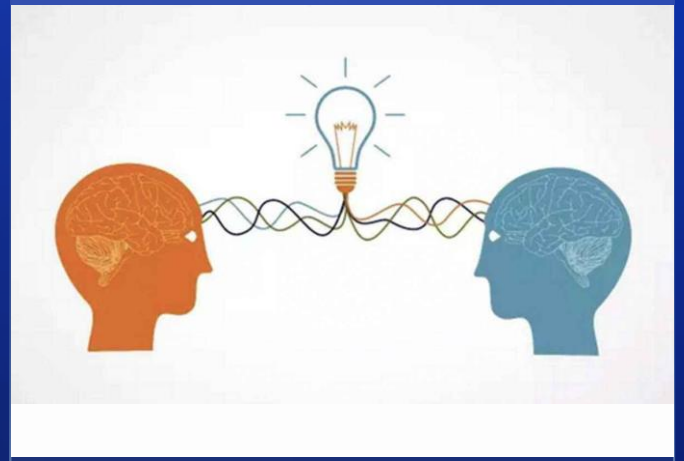
- 质控数据分析与监督

CONTENTS
目录

以ChatGPT为代表的大模型技术特点



更强的学习能力



更好的迁移能力

更深的理解能力

国产大模型现状

技术原理没有壁垒，落地能力和客户价值成为关键

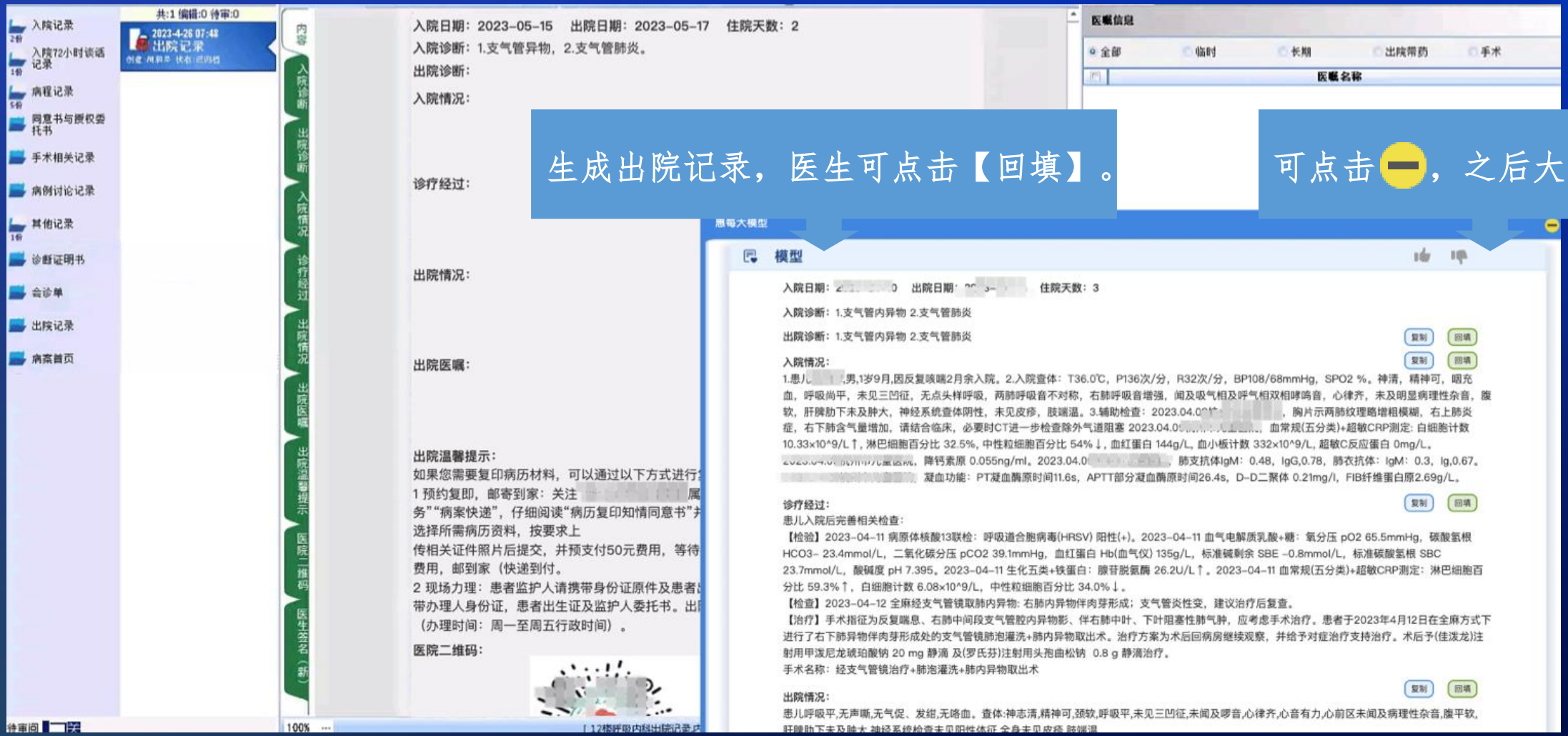
医疗行业的特殊性要求低成本实现大模型私有化部署

医院数据安全特殊性
内网使用
患者数据不能出院

世界级难题
全球GPU芯片紧缺

医疗大模型应用场景：病历生成

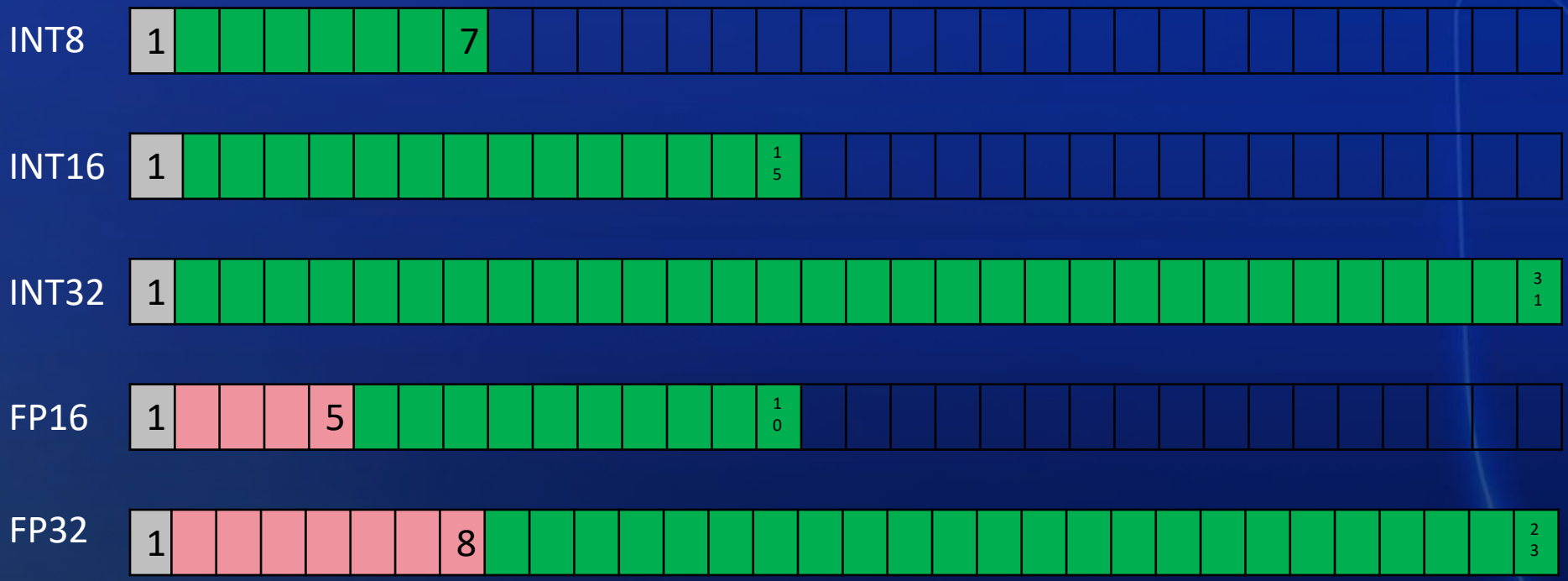
出院小结需要对多类数据进行总结并形成摘要，契合大模型技术本身的优势



CONTENTS
目录

加速方案 –模型量化

降低模型参数的精度来实现加速



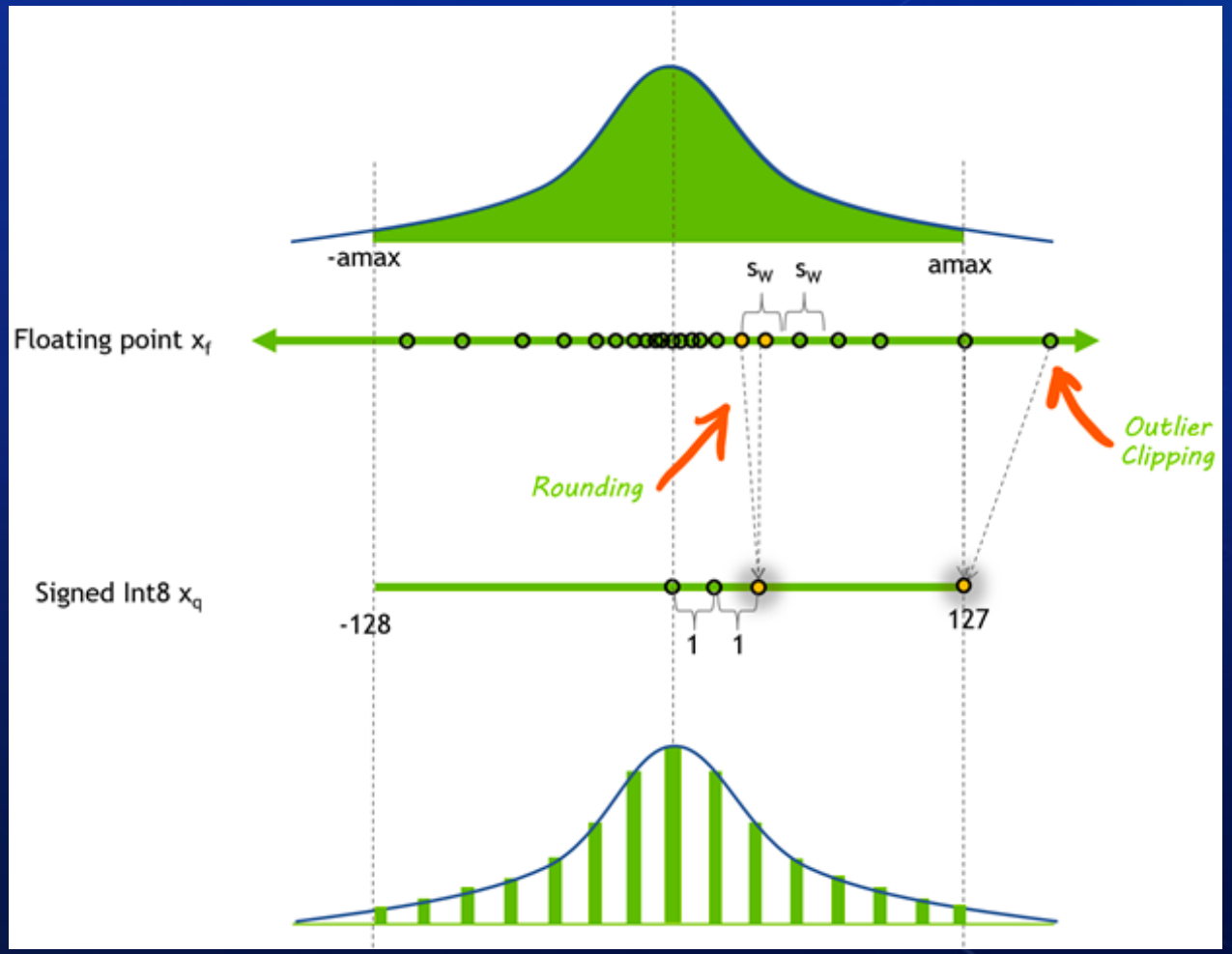
量化方案 - 量化算法

在线量化(On Quantization）

在特定的操作内嵌入伪量化节点(fake quantization op)，用于统计训练时流经该节点数据的最大值和最小值，便于在推理时进行量化，而且训练时量化指干预神经网络的前向推理，不参与反向传播和梯度更新。

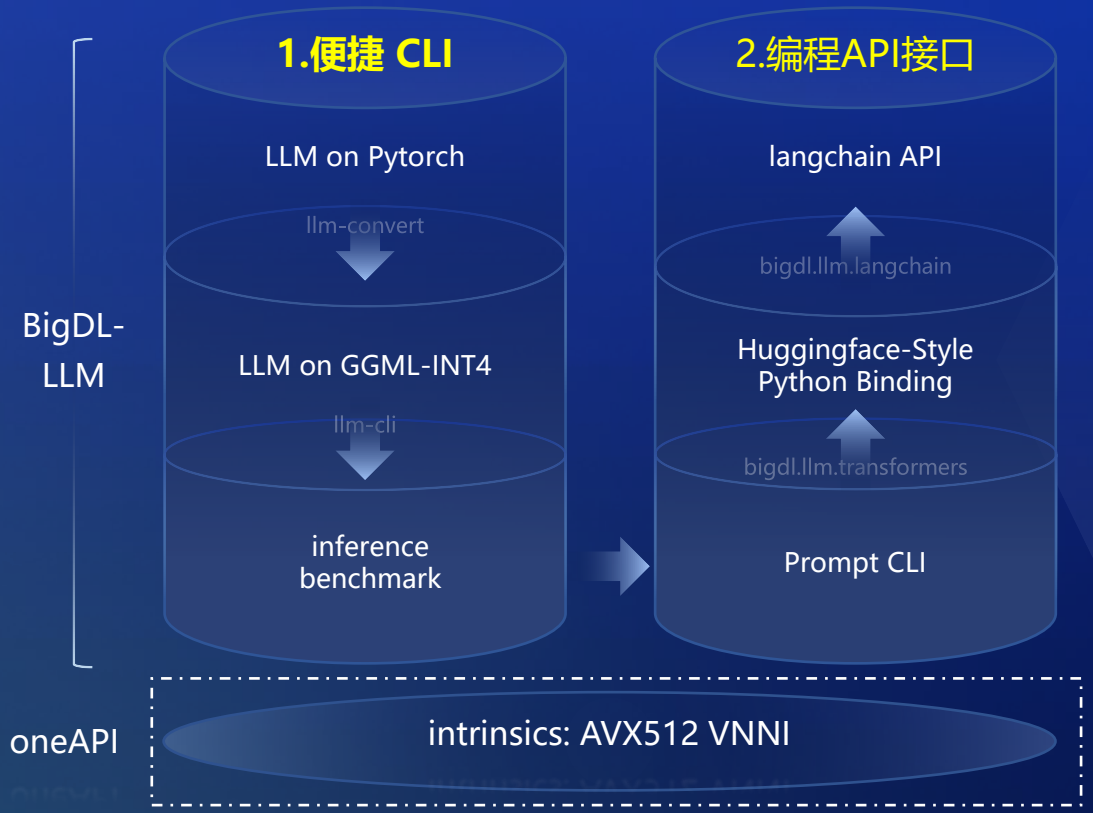
离线量化 (Off Quantization)

基于依据训练完的模型进行量化，工作原理比较简单，对网络模型的权重参数通过量化公式进行转换，但在内存初始化时对网络模型的权重进行反量化操作变成float进行正常的推理。



量化方案 - 基于CPU指令集上的BigDL-LLM优化方案

基于第四代英特尔® 至强® 可扩展处理器



1. 便捷 CLI (预览和快速测试)

```
llm-convert \
-x <LLM家系名> \
-f pth \
-t <量化精度> \
-o <输出目录> \
<模型源目录>
```

```
llm-cli \
-t <线程数> \
-x <LLM家系名> \
-m <量化模型文件> \
-c <上下文长度> \
-n <最大输出长度> \
-v -p \
"<Prompt>"
```

2. 编程 API接口 (代码整合)

HuggingFace API

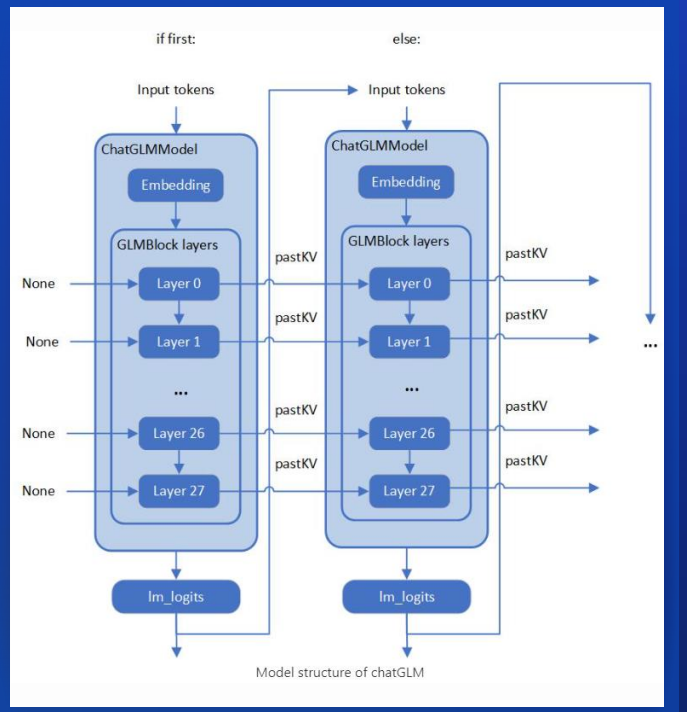
```python
from bigdl.llm.transformers import AutoModelForCausalLM
model = AutoModelForCausalLM.from_pretrained('/path/to/model/', load_in_4bit=True)
```

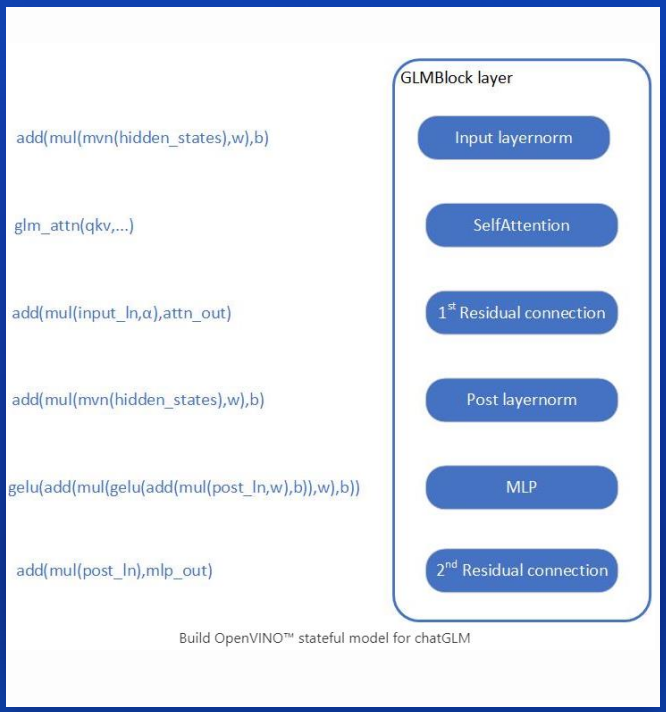
Langchain API

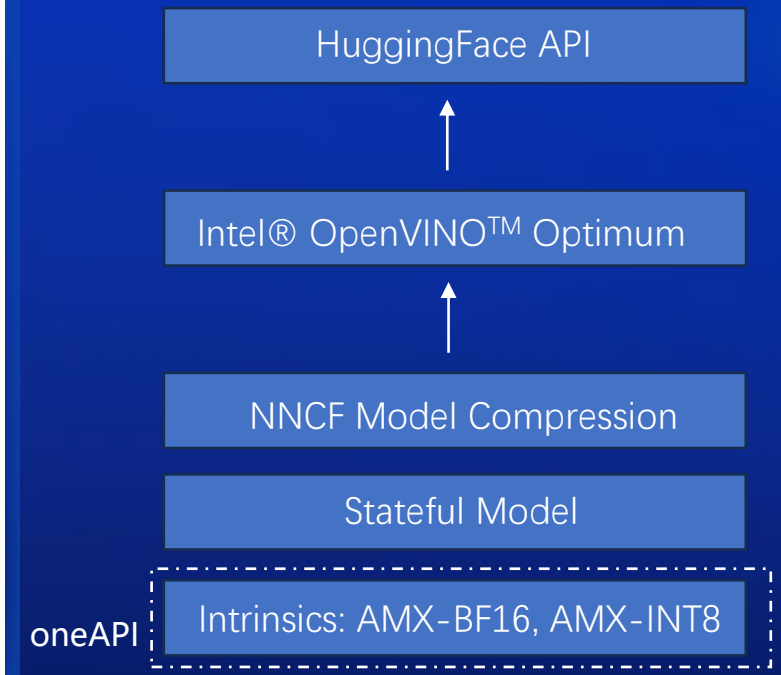
```python
from bigdl.llm.langchain.llms import TransformersLLM
from bigdl.llm.langchain.embeddings import TransformersEmbeddings
from langchain.chains.question_answering import load_qa_chain

embeddings = TransformersEmbeddings.from_model_id(model_id=model_path)
bigdl_llm = TransformersLLM.from_model_id(model_id=model_path, ...)

doc_chain = load_qa_chain(bigdl_llm, ...)
output = doc_chain.run(...)
```

https://github.com/intel-analytics/BigDL/tree/main/python/llm

非量化方案 - 键值缓存、算子融合的OpenVINO™加速方案







1 利用零拷贝视图传递KV，实现KV缓存加速

2 引入AMX加速BF16/INT8计算，并实现算子融合

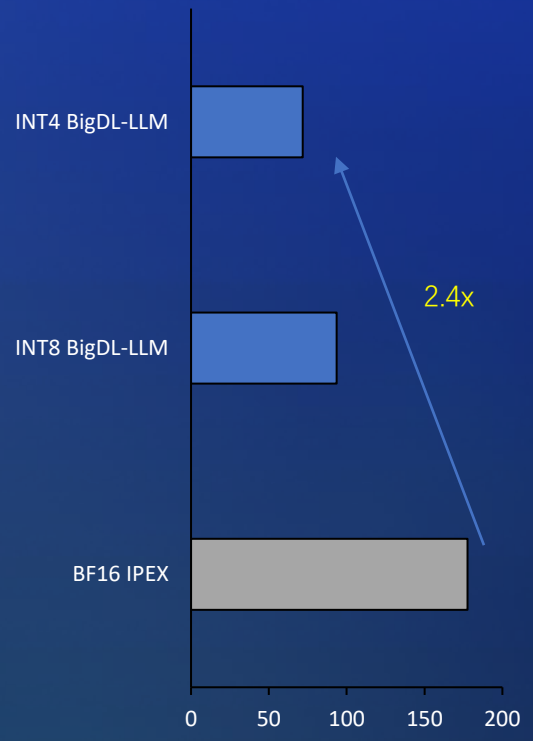
3 使用 Optimum 让优化效果可扩展

https://blog.openvino.ai/blog-posts/enable-chatglm-by-creating-openvino-tm-stateful-model-and-runtime-pipeline
https://huggingface.co/docs/optimum/main/intel/inference

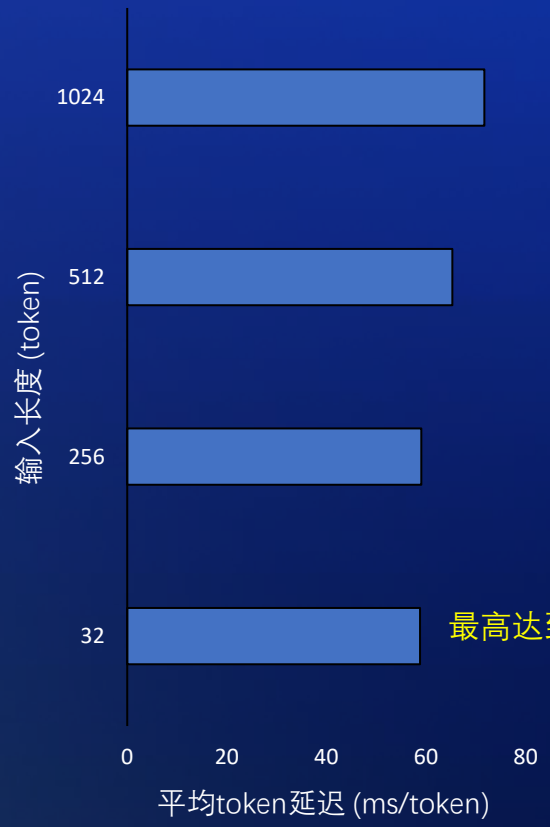
医疗大模型CPU推理技术

第四代英特尔®至强®可扩展处理器
英特尔®BigDL-LLM量化方案
加速阿里云c8i-2xlarge上推理(8核)

输入为1K时不同精度下的next-token延迟



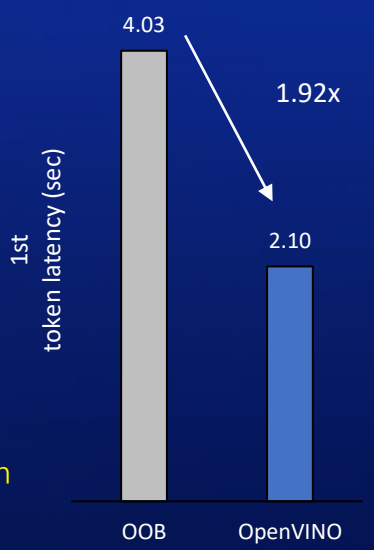
INT4 BigDL-LLM在不同输入下的next-token延迟



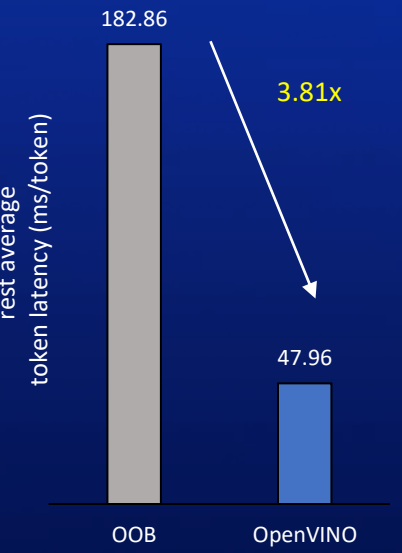
最高达到58.8ms/token

第四代英特尔®至强®MAX 处理器
英特尔®OpenVINO™-LLM非量化方案
加速英特尔®至强®MAX实例加速性能(48核)

首词延时下降
(2K输入)



其余平均延时下降
(2K输入)



平均token延迟48ms/token

- 1S-SPR-HBM Quadrant/HBM-Cache

CONTENTS
目录

医疗大模型CPU推理技术实践总结

- 惠每科技基于医疗大语言模型实现了CDSS业务上的创新
- 基于英特尔®SPR-SP和SPR-HBM实例，加速实现了惠每科技CDSS业务上的性能提升
- 推荐配置：

在搭载了第四代英特尔® 至强® 可扩展处理器(代号SPR)的实例上使用量化加速方案；

在搭载了英特尔®SPR-HBM的实例上使用非量化加速方案

惠每科技 | Dr.Mayson
Clinical Decision Support System

谢谢！