

AI智能体轻量化技术栈

模型层优化

量化

FP32→INT8/INT4

剪枝

移除冗余连接

知识蒸馏

大模型→小模型

架构搜索

自动优化结构

软件层优化

推理引擎

ONNX/TFLite

算子融合

合并计算操作

内存优化

复用/分片加载

并行计算

多核/批处理

硬件层加速

专用AI芯片

NPU/TPU/VPU

异构计算

CPU+GPU+NPU

指令优化

NEON/AVX/SIMD

功耗管理

动态调频/休眠

系统集成与协同优化

模型-软件-硬件协同设计，实现最优的性能、功耗和成本平衡