OPPO 多模态大模型端侧化应用实践

演讲人：宋晓辉

OPPO 端侧化算法组负责人

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

01

端侧化算法技术概览

模型端侧落地的困难

珍贵的内存

终端设备内存有限，算法效果需要一定程度上向**模型体积**和**推理精度**妥协，因此如何设计和优化端侧**模型压缩**算法，利用**有限的内存占用**获得**最佳的算法效果**，为用户提供流畅、好用的端侧AI功能，是端侧化算法持续追求的目标之一。

有限的电量

为用户提供端侧AI能力的同时，也不能成为“**电老虎**”，不能制造续航焦虑，持续的下探端侧AI的能耗水平，需要端侧算法、工程和芯片团队的共同努力。

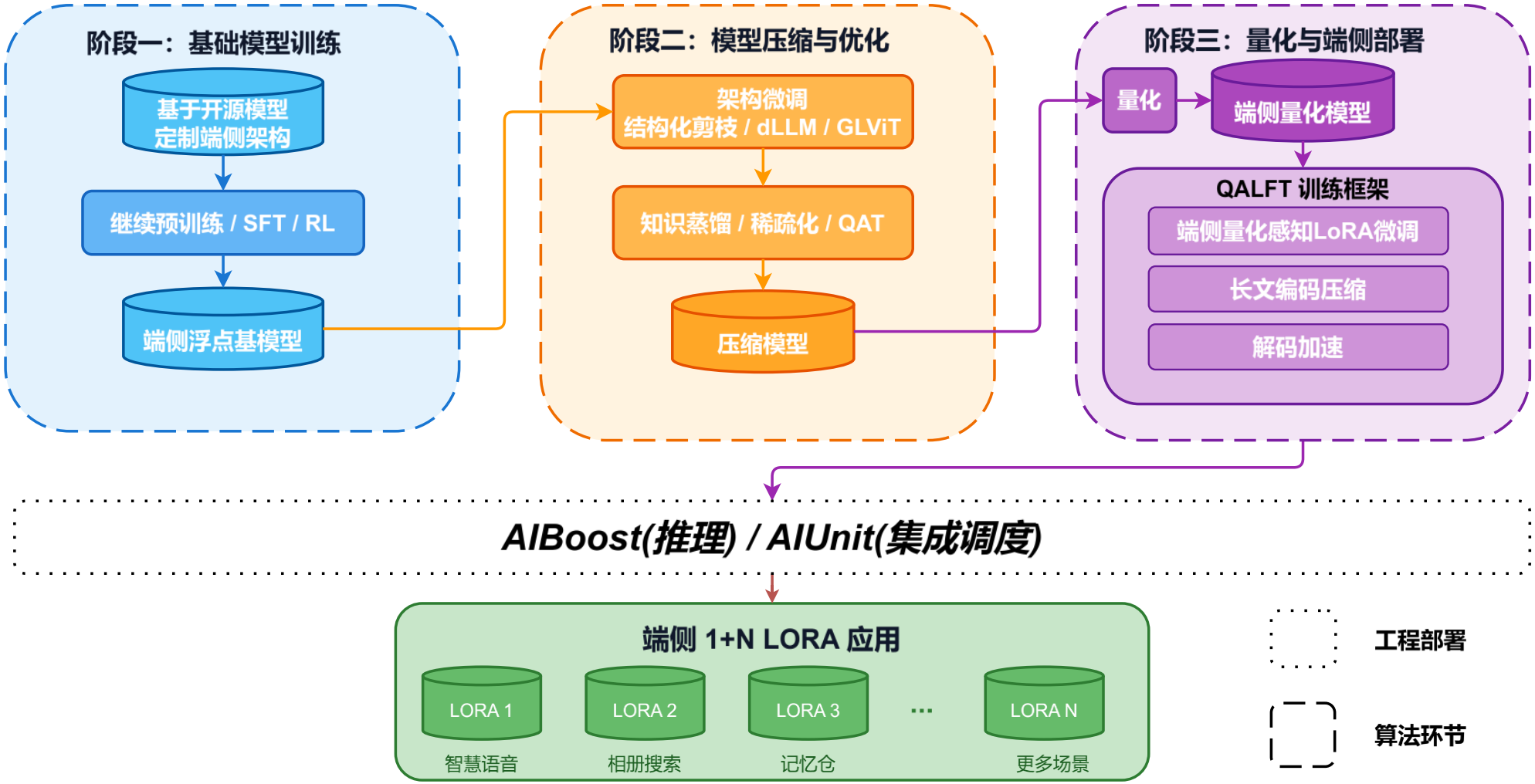
强耦合的业务

为了减少对用户手机ROM空间的占用，所有端侧算法能力共用同一个基模型，因此基模型迭代需要和众多业务需要保持相同的节奏，这对**工程进度管理**，算法的迭代和测试验收都是很大的考验。

较高的OTA成本

端侧模型体积大，不能高频OTA更新，同时缺乏运营日志，难以敏捷迭代修复问题。因此端侧业务的交付，需要在**算法质量，稳定性，性能功耗热**等多个维度进行严格的测试。

模型端侧化算法技术概览



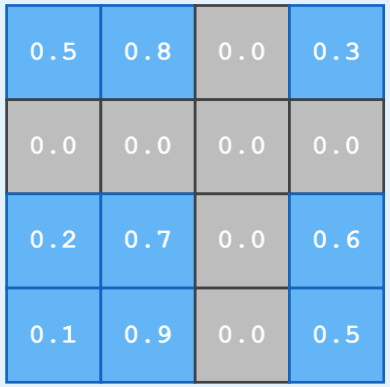
02

模型稀疏化压缩

模型稀疏化

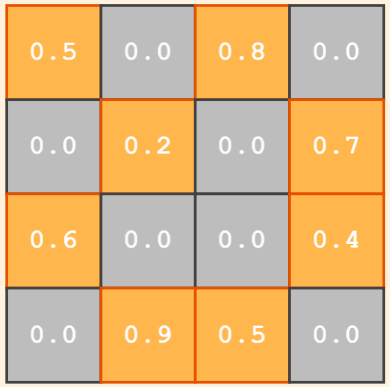
模型稀疏化三种范式

结构化稀疏（剪枝）



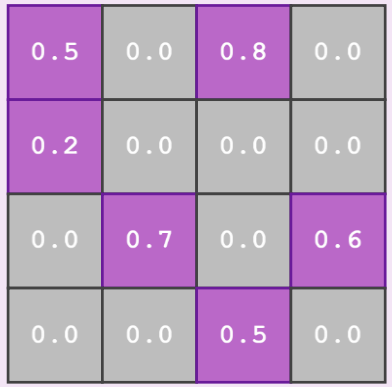
整行/整列置零
通过删除整行或整列参数
减少模型层的维度
实际完成了结构剪枝操作
可直接减少计算量和存储

N:M 结构化稀疏



2:4 结构化稀疏
每连续4个元素中置零2个
利用 NVIDIA Tensor Core
专门硬件加速支持
可获得计算加速效果

非结构化稀疏



不规则稀疏模式
通过硬件编码减少权重存储大小
降低 Memory Bound 瓶颈
提升解码速度与降低功耗

结构化稀疏–L0正则化方法

核心思想

为每个参数学习一个可微分的「门控值」(0~1)，训练中自动识别重要参数（门控→1）和冗余参数（门控→0），实现端到端的稀疏化学习

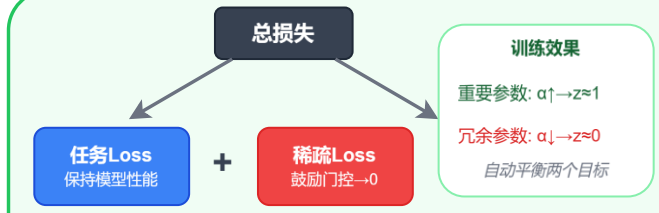
1 初始化门控参数

初始时所有参数都保留，门控值=1表示完全激活

2 训练中：噪声采样 + 动态门控

关键：噪声使门控可微分，同时保持随机性探索

3 通过Loss学习重要性



关键：剪枝 Loss控制稀疏度，任务Loss保护重要参数

结构化剪枝：整行/整列门控值演化

图例：门控≈1 (保留)　门控中间态　门控≈0 (剪枝)　待删除行/列

结构化稀疏–L0正则化方法

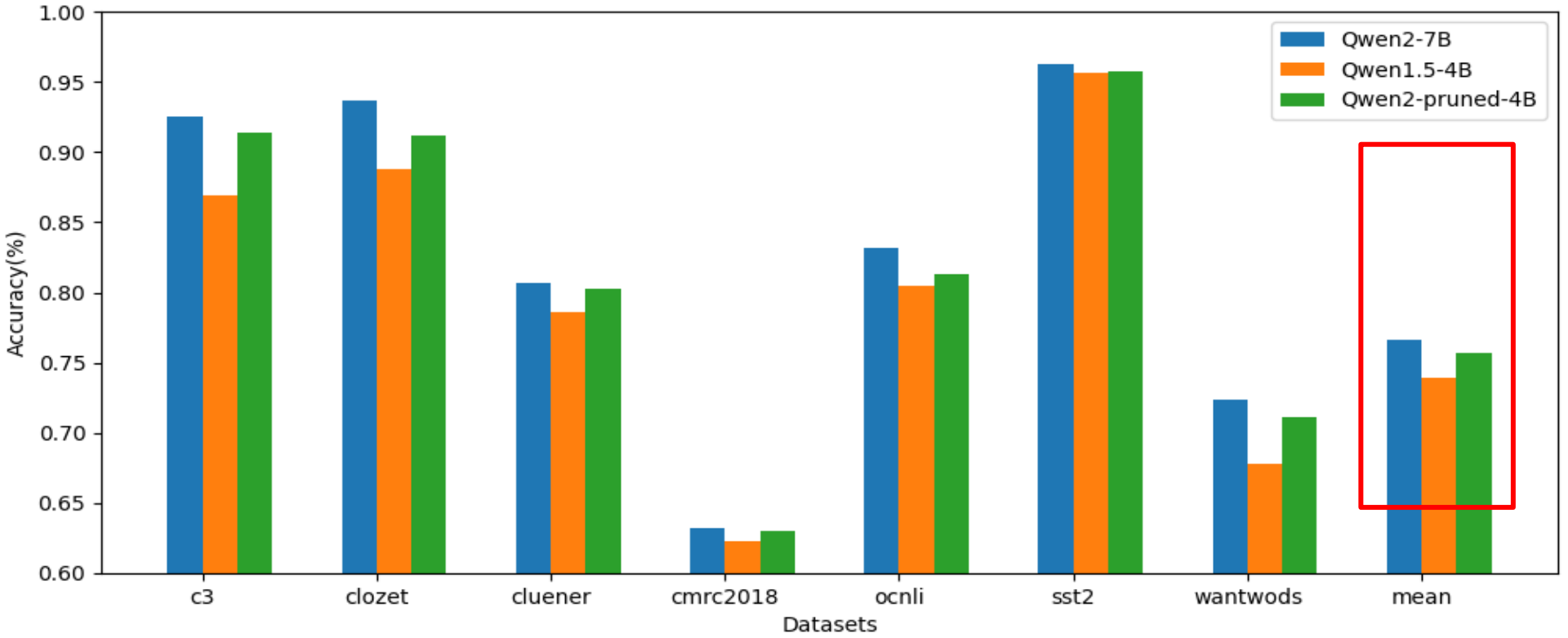
缺点	优化方向	核心优化思路	具体改进措施
收敛速度慢，对数据规模要求比较高，mask的优化速度和模型权重的优化速度不好均衡	如何让lm_loss通过mask对权重的评估更快、更准确	把L0正则化剪枝直接通过梯度下降获得0–1二值mask转换为两阶段问题： • **排序**：通过梯度下降评估参数重要程度，体现在mask数值的排序上 • **剪枝**：通过soft top–k mask将排序结果渐进的转化为0–1二值序列	加入**梯度缩放因子**，将mask从0–1映射到更小的范围，e.g. (0,1e–3]，并通过**伪输入**技巧将mask引入到lm_loss的优化过程中，让mask对模型效果感知更强，提升优化效率。
均匀分布的噪声均值太高，噪声波动影响的参数量较大（类似dropout），无法适用于剪枝比例比较大的情况(例如90%以上)	Mask的噪声分布设计和实现		使用加入**直通估计**的hardtanh，更加充分的利用mask的梯度信息。 重写**噪声采样逻辑**，转变噪声分布，并限制噪声的绝对值的上限，仅用于评估重要性，不致力于产生二值mask。
为了达成剪枝目标，会产生较多0–1之间的mask，剩余参数存在浪费，影响剪枝后模型的效果	概率累积分布函数的设计和实现		整合了Hard Concrete Distribution和soft–topk–mask的思路，设计了一些列辅助函数，保证剪枝目标达成并且不浪费参数。

排序和剪枝两个阶段在训练过程中动态交替进行，实现了较为平缓的剪枝过程。

结构化稀疏–落地实践

基于OPPO的L0改进算法，从Qwen2–7B剪枝到4B，超过了Qwen1.5–4B的效果。

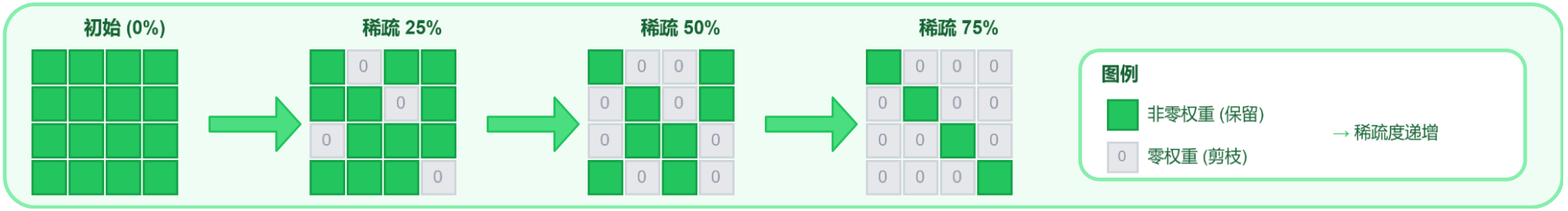
L0正则化剪枝在Qwen7B->4B的结果



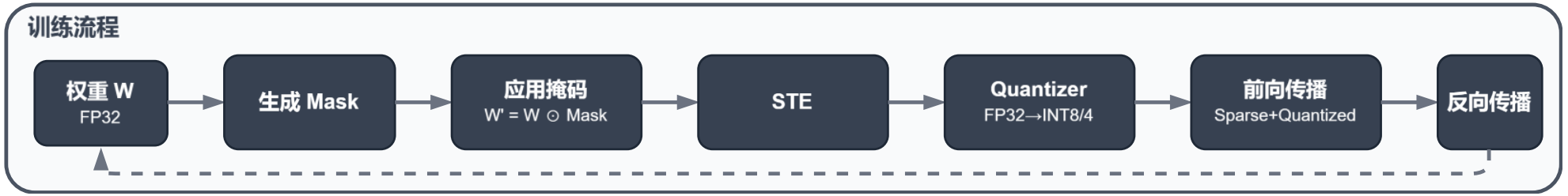
在ColorOS 15.0的端侧基模型剪枝和解码加速的draft model上都有应用。**支撑OPPO智慧语音端侧化业务**。

非结构化稀疏（内存压缩）

权重稀疏化过程



稀疏化训练过程

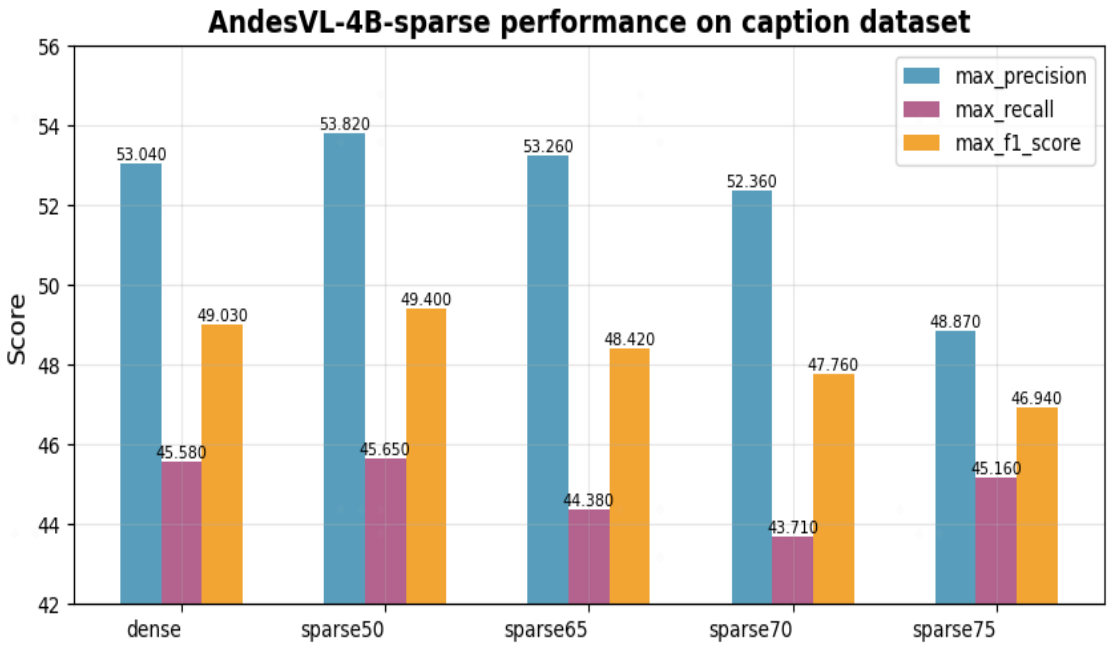


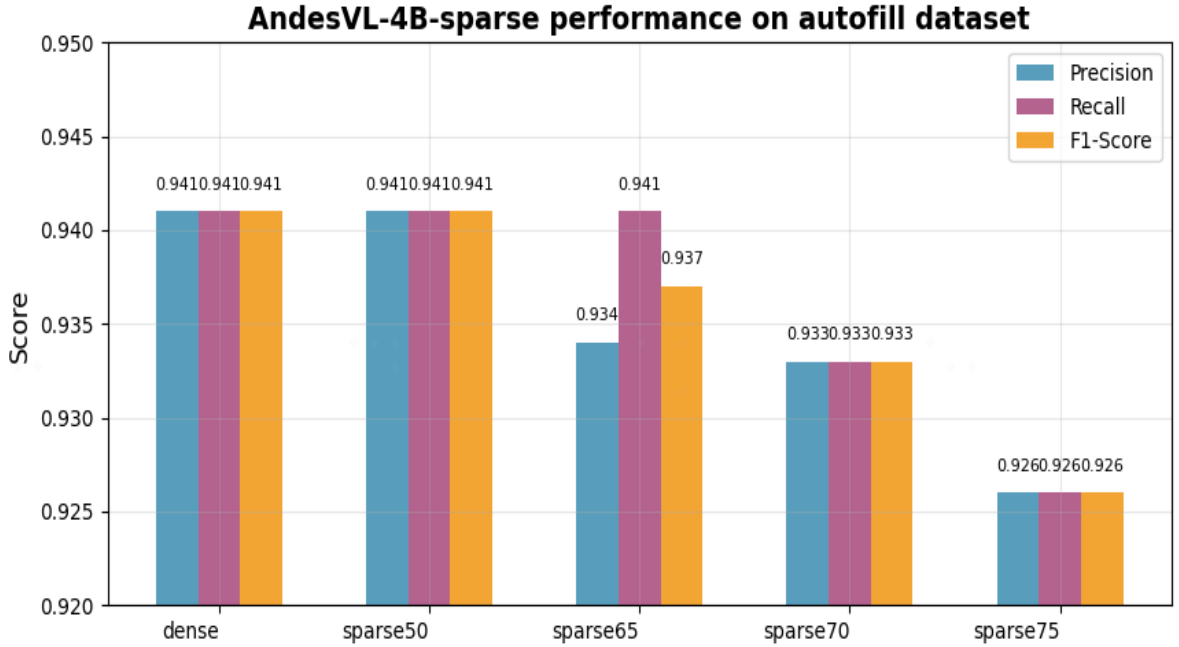
稀疏化训练策略

非结构化稀疏–算法效果

在AndesVL–4B 模型分别进行了 50%、65%、70% 和 75% 四个稀疏度的模型训练，并在image caption和多模态信息抽取任务上进行效果验证：

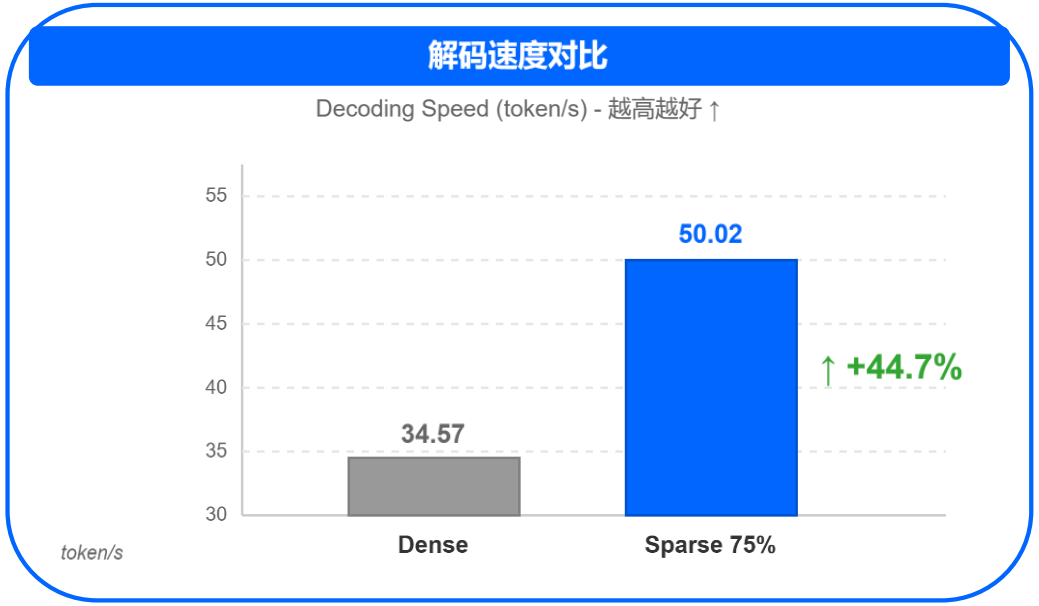
1. **50% 稀疏度**模型在两项测试任务中表现优异，其综合性能指标与原始稠密模型**基本持平**，甚至**展现出轻微的性能优势**，表明适度的稀疏化可能带来正则化效应。
2. 随着稀疏度提升至 65%–75%，模型性能呈现可控范围内的温和下降，性能衰减曲线显示：稀疏度每增加 15%，性能损失约 2–3 个百分点，**即使在 75% 的较高稀疏度下，模型仍保持核心能力，各项关键指标的下滑幅度均小于 5%**，为模型部署提供了显著的效率提升空间。

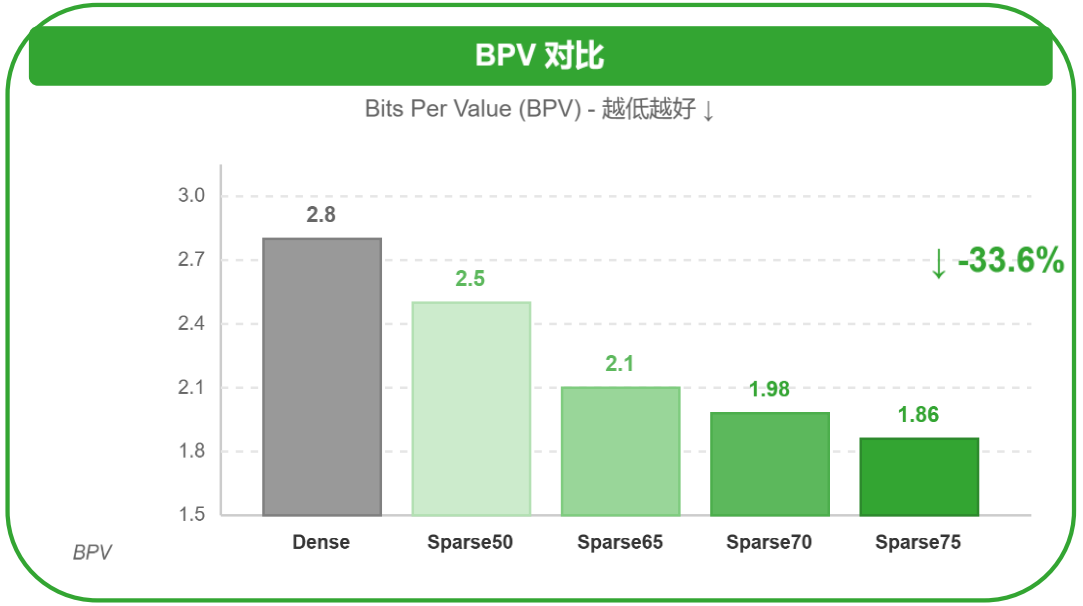




非结构化稀疏–性能收益

在MediaTek天玑9500芯片上，结合硬件压缩能力，实现速度和内存的大幅度优化。





结论：75% 稀疏度下，解码速度提升 44.7%，BPV 降低 33.6%，有效实现模型压缩与推理加速的双重优化。

03 量化感知训练

端侧基模型量化感知训练

OPPO自研**量化感知训练框架**，支持以下feature

1. 高度灵活的点位设置和quantizer模式
可以快速适配新模型，支持权重、激活的per tensor/channel/token/group的asym/sym的fakequantizer设置以及动态、静态的QAT训练模式。

2. 动态精度分配策略
可细粒度识别高敏感权重和激活点位，实现自动化的混合精度QAT。

3. 端到端静态量化
与工程团队、片商紧密合作，通过点位对齐和静态QAT训练，可以将QAT得到的量化编码直接导入到端侧模型中，**绕过PTQ过程**，减少算法迭代过程中的效果不确定性，通过这种方式可以支持低bit量化的高效部署（W2，A8等）。

Model	DocVQA (test)	InfoVQA (test)	TextVQA (val)	ChartQA (test)	Overall
AndesVL-4B-Instruct-Base (PTQ)	93.2	89.0	91.4	89.3	90.7
AndesVL-4B-Instruct-Base (QAT+PTQ)	95.4	95.2	97.5	95.1	95.8

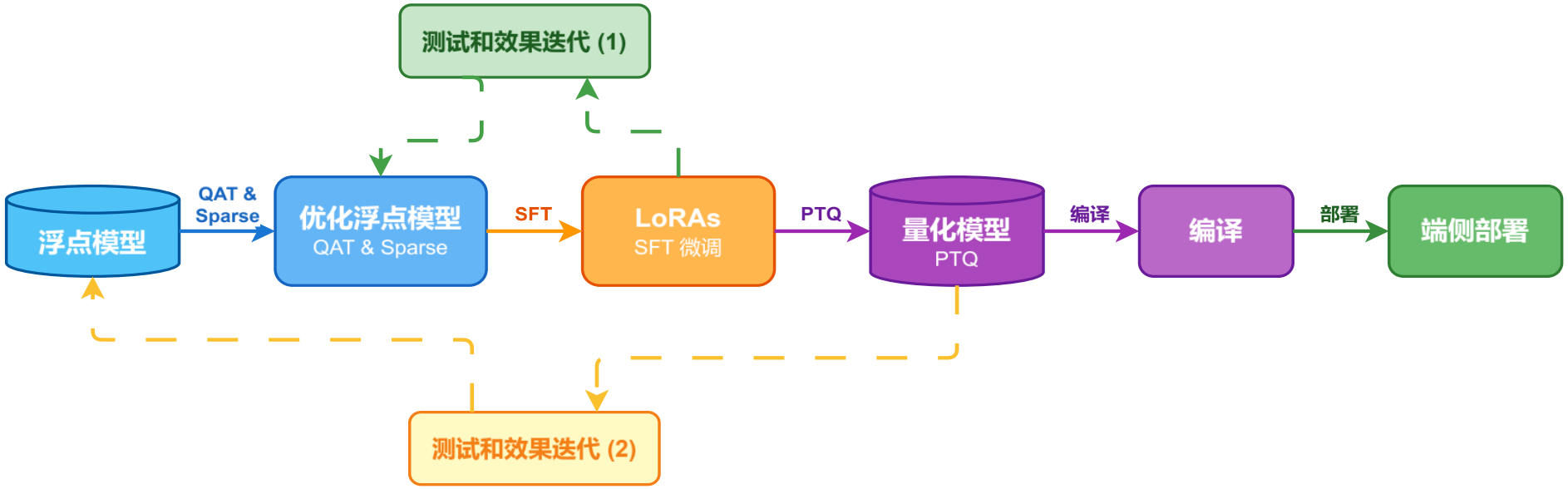
Table 17: Top-1 overlap between AndesVL-4B-Instruct-Base (PTQ) and AndesVL-4B-Instruct-Base (QAT+PTQ) on 4 OCR benchmarks.

通过QAT训练，可以将模型端侧和浮点的输出一致性(Top1–Accuracy)保持在0.95以上。

垂域场景落地–QALFT训练框架

纯浮点模型落地过程中的问题：

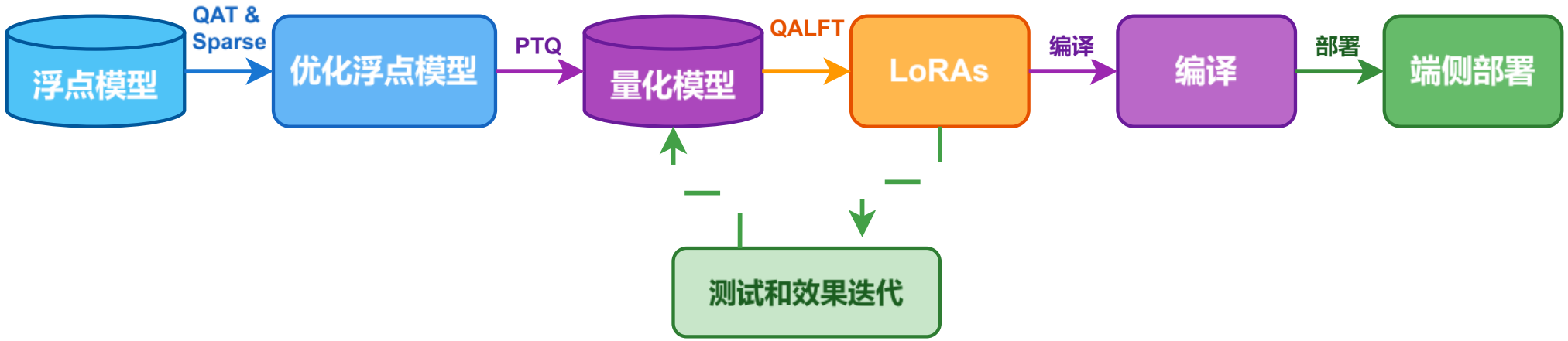
- **PTQ损失不可控：**浮点lora优化完成之后，各个业务一起做量化时对彼此的影响不可控制。
- **业务间耦合过重：**每次LoRA更新时，必须对基础模型及所有LoRA进行重新量化，产生显著时间成本。
- **测试资源投入较大：**从浮点模型到端侧落地需经历多环节协同。但是如下图所示，一旦效果不达预期，需要重新迭代整个流程。



QALFT训练框架

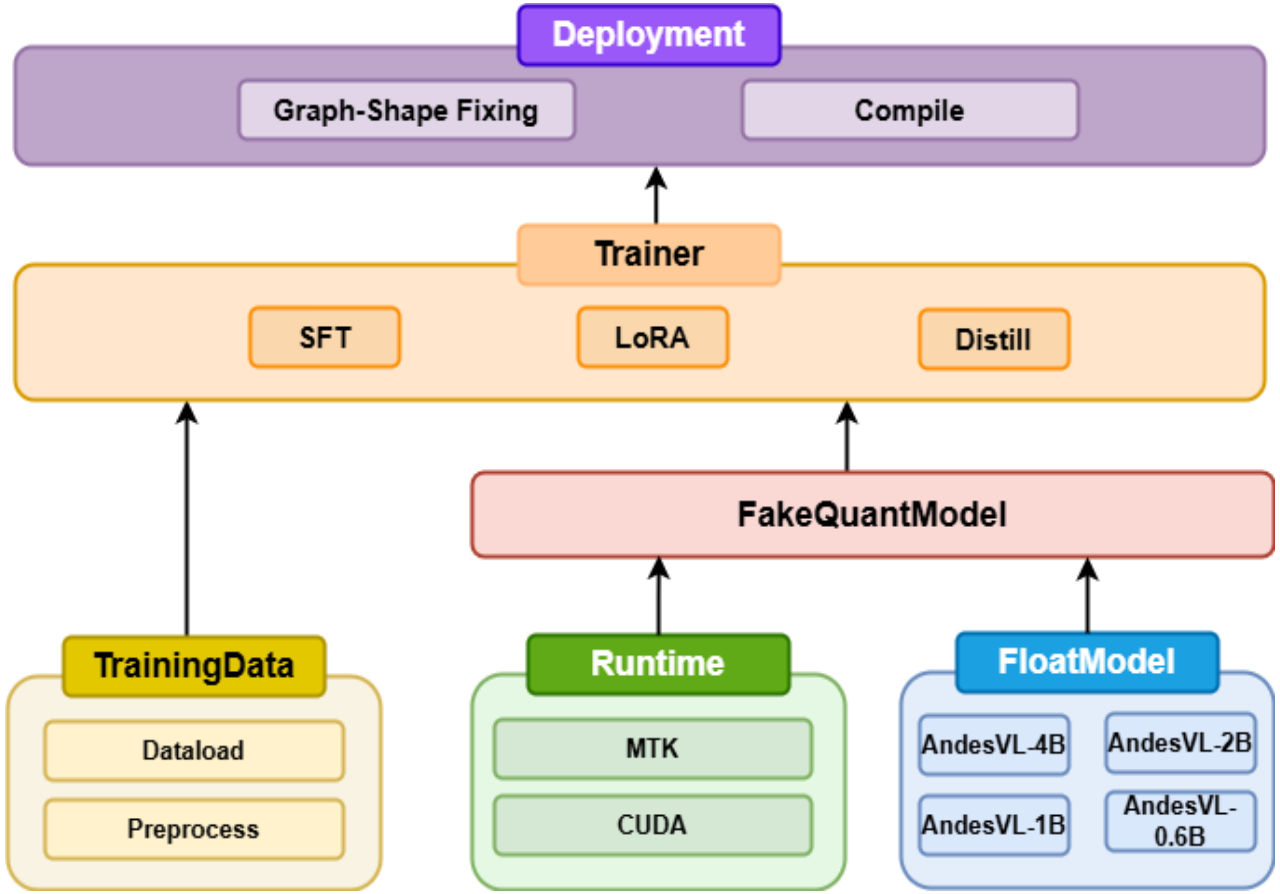
为解决上述问题，我们开发了**量化感知LoRA微调框架**（QALFT）。用户能够使用该工具直接在量化后的模型上进行微调，无需关注模型量化的具体技术细节。这一方式将模型训练与部署流程简化为两个核心步骤：

1. 针对业务场景，基于量化模型进行微调。
2. 评估量化模型在业务场景的实际效果。



QALFT训练框架

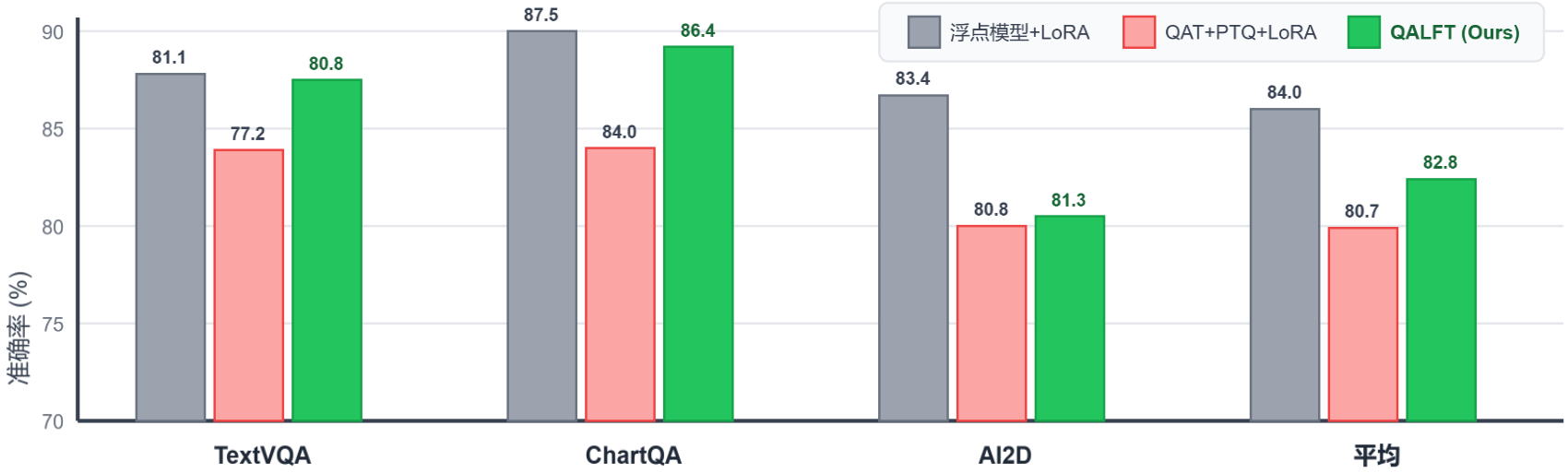
- **模块解耦架构：** 模型、数据、训练器三方解耦，各组件可在不同训练流程中复用。
- **丰富的训练范式：** 适配SFT / 蒸馏等多种训练范式。
- **扩展灵活性：** 支持新模型、数据格式、训练方法的可插拔接入。
- **底层平台隔离：**MTK/Qualcomm平台库与数据、训练器等上层设施隔离。
- **零代码化部署：** 无需额外代码开发，支持微调、验证全流程操作。



QALFT训练框架

使用QALFT具有以下优点：

- **降低人力成本，缩短部署周期：** 显著减少了传统方案中LoRA相互不独立以及两次效果迭代所需的人力与时间投入。
- **更适配1+N LoRA模式：** 流程与“1+N LoRA”模式更加匹配，即量化的基模型保持通用性和业务无关性，并通过LoRA的方式对业务场景进行适配。
- **更好的算法精度：**片商提供和端侧推理一致的模型量化推理库，实测算法效果显著优于PTQ方案（**端侧效果相比浮点微调损失能控制在3%以内。**）



平均效果损失：QAT+PTQ: -3.93% vs **QALFT: -1.43%**

04 编解码加速

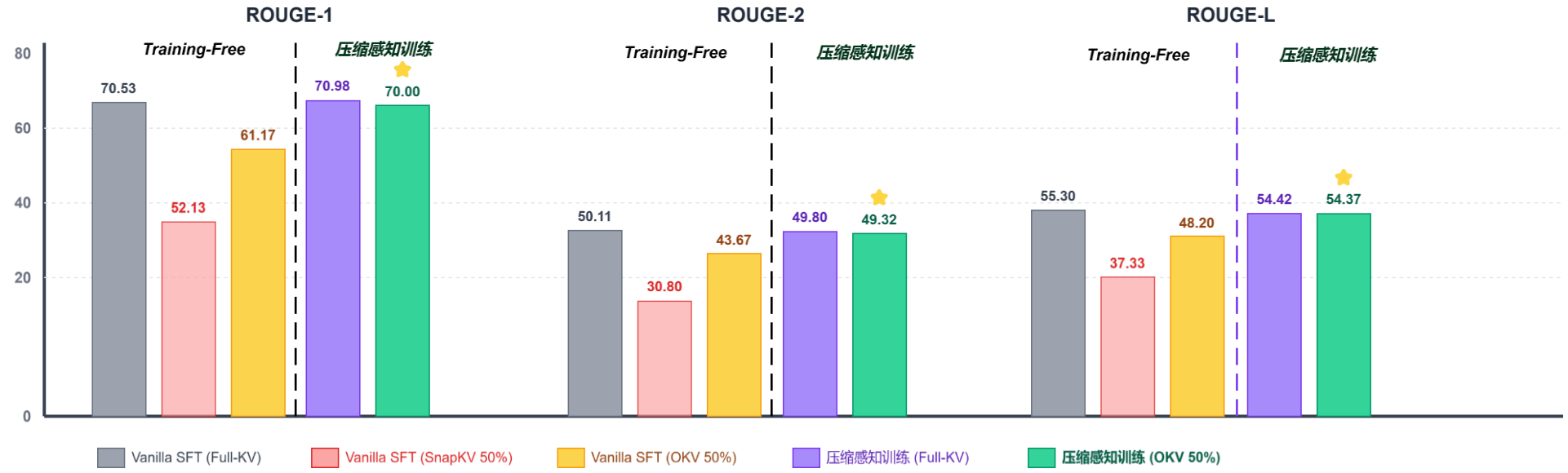
KV-Cache压缩&KV-Cache压缩感知训练

端侧长上下文存在的挑战

内存：随着输入长度增加，KV缓存呈线性膨胀，内存占用迅速提升。

计算：由于端侧算力有限，长文本下softmax等计算量急剧膨胀、推理延迟显著增加，制约了长上下文能力的广泛应用。

自研training-free压缩方案&压缩感知训练方案在通话摘要场景的对比



压缩感知训练方案能显著降低cache-eviction算法损失。

解码加速（ speculative decoding）

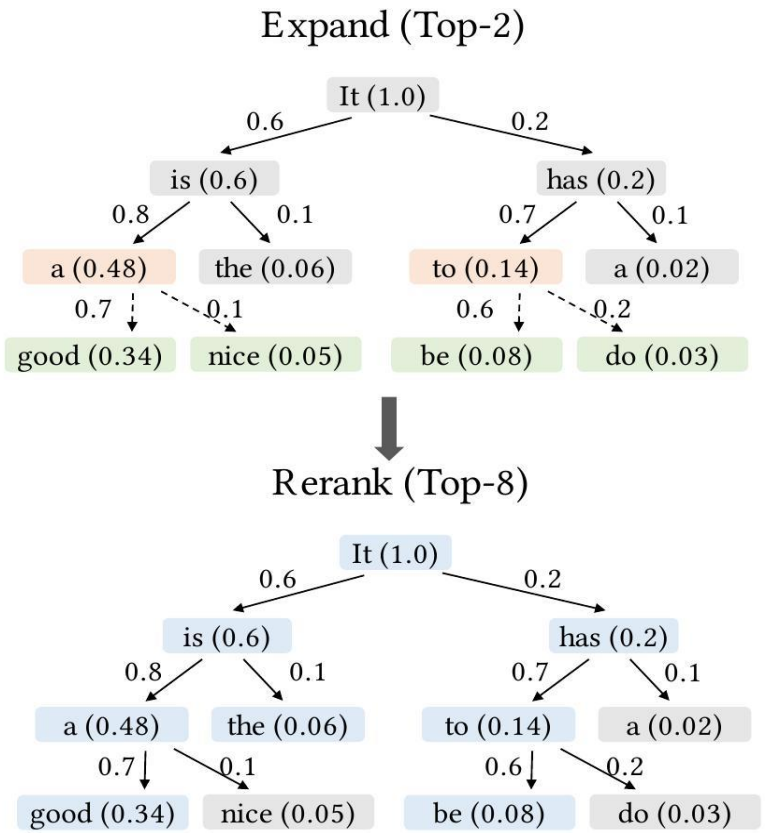
3次推理
接受1+3+3=7个token。

解码加速的核心优化问题：

1. **提升草稿的接受率和接受长度。**
2. **降低草稿模型的执行成本。**

解码加速

采样树(topk & total tokens)：

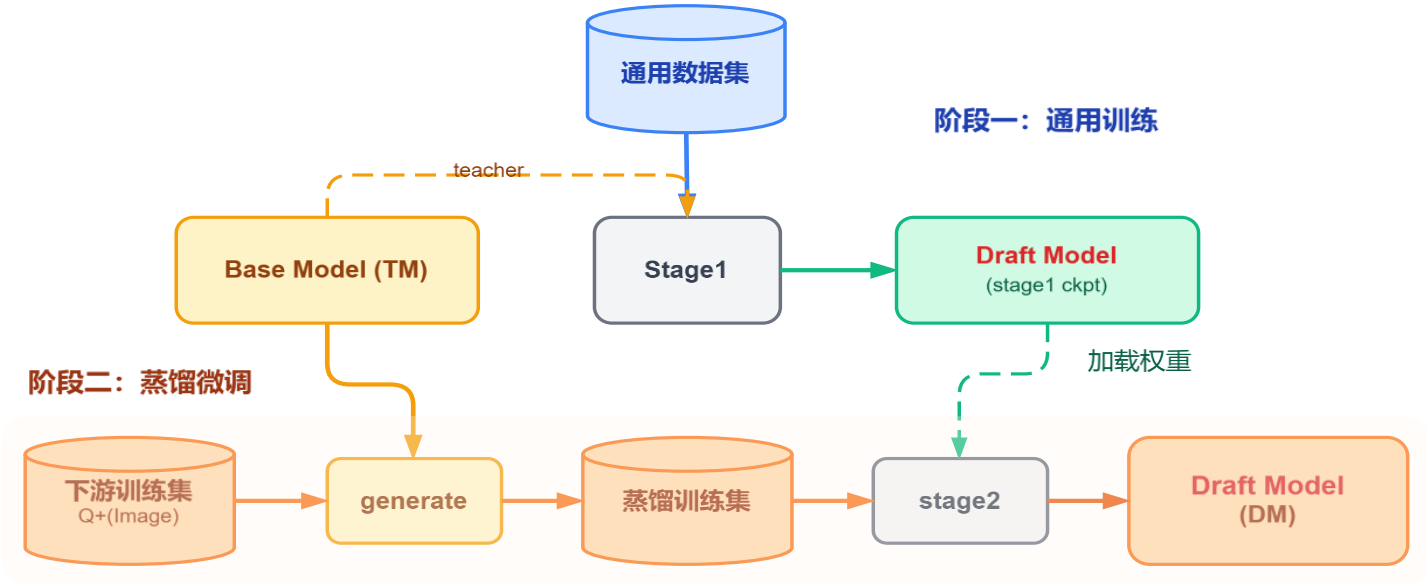


1. topk
   - 每次调用小模型时，都只选出topk个词往后扩展，避免层数过大爆炸。
   - 草稿模型的ARN ⩾ topk

2. total_tokens
   - 用于最后排序出得分最高的total_tokens个路径进行一次并行验证。
   - 目标模型的ARN ⩾ total_tokens

解码加速

训练流程



推理优化

路径选择、路径扩展策略优化；早停机制。

业务效果

多模态理解场景：

- AndesVL2-4B(aimv2+qwen3-4B):通过算法和工程侧的联合优化，得到**6.7x**的端到端加速效果（23.3 t/s -> 157 t/s）
- AndesVL-4B(internvit+qwen2.5-3B)，峰值速度达到**240+tok/s**

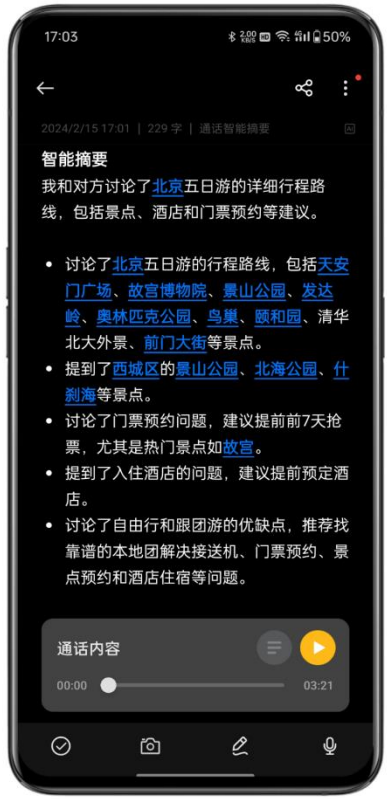
05

落地实践

业务落地

智慧语音端侧化

- OS14.0 **行业首个**端侧7B大模型落地
- OS15.0 自研剪枝蒸馏算法，**升级1+N lora架构**，支持通话、三方应用、视频摘要和实体组合。
- OS16.0 **升级端侧多模态大模型**基座，实现更多场景基模型复用，极致的模型压缩和解码加速，**大幅度降低模型运行内存功耗**。



AI搜索、记忆仓（多模态能力）

- 提供端侧相册caption能力，支撑AI搜索业务。
- 提供端侧相册卡证信息抽取能力，支撑记忆仓-自动填充业务。

技术预研

融合芯片级上下文加速技术

(Prompt 压缩技术和 kv cache eviction技术)

端侧支持处理

128K 超长上下文

可本地处理

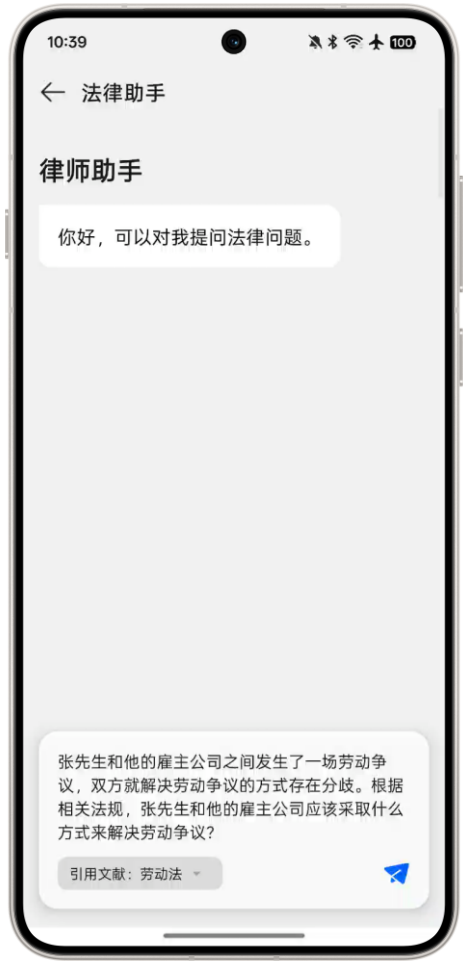
20 万字级文档

(如 300 页书籍)

赋能法律、医疗、教育等专业场景分析

突破端侧能力边界

端侧超长上下文支持



技术预研

极致模型内存压缩技术结合先进解码加速技术，
通用场景峰值推理速度达

240 t/s

高效办公快人一步

端侧极速生文

技术预研

AI 一键闪记端侧版，全场景啥都能记

端侧高效多模态信息处理

4s 完成多次 4B 模型推理

响应速度不输云侧，无网也能随心记

端侧一键闪记

06 总结和展望

总结与展望

可能用在哪里

- 端云协同的agent应用
- 隐私和个性化
- 新硬件、具身智能
- 开放的端侧算力算法服务

技术上还需要优化什么

- 面向NPU的高性能架构
- 模型量化压缩、推理加速
- OS层面的资源调度、模型更新
- 成熟稳定的开发工具链

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会