CANN架构下的招投标领域大模型长文本推理优化实践

面向国产NPU的大模型推理工程实践与优化策略

分享嘉宾：熊文韬

2025.12.6　　中国·深圳

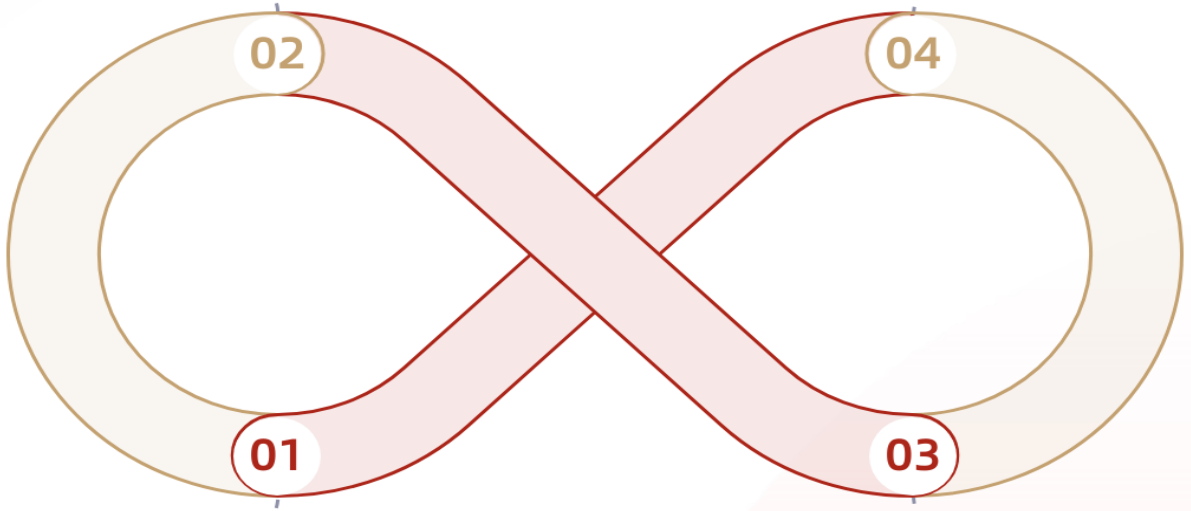
content

目录

背景：政策导向与算力迁移趋势

01

政策驱动：算力基础设施的战略转型



时代机遇

要求关键基础设施
完成替代

监管目标

供应链安全考量推
动算力自主化进程

算力突破

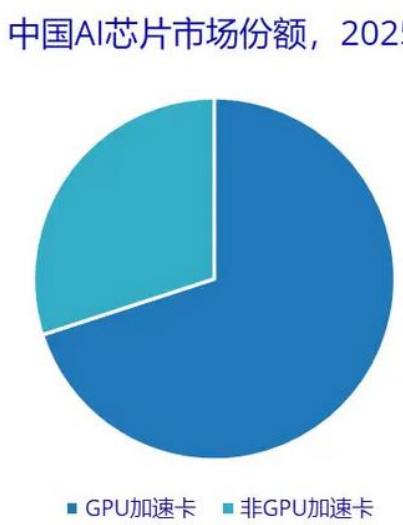
华为昇腾系列NPU
成为产业首选方案

技术升级

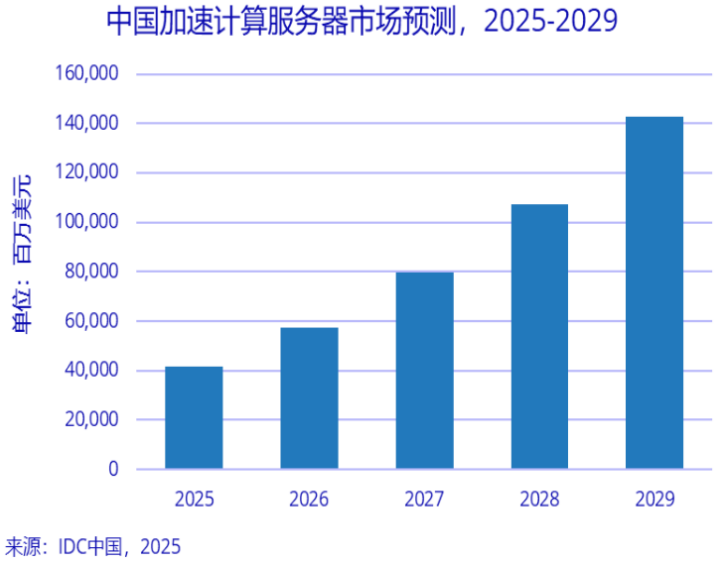
CANN软件栈提供完
整技术软件栈

CANN

市场格局：非GPU算力快速崛起



来源：IDC中国，2025



来源：IDC中国，2025

30%

非GPU市场份额

2025年上半年NPU与CPU加速卡需求增速远超GPU

35%

国产芯片占比

国产AI芯片整体市场份额持续上升

1400+

优化算子库

CANN内置深度优化的算子数量

数据来源：IDC市场统计报告（2025年上半年）

CANN

大模型浪潮：算力需求的指数级增长

1

通用模型时代

ChatGPT等大规模预训练模型出现,参数量突破千亿级

2

垂直领域深化

行业专用大模型涌现,上下文长度快速扩展

3

国产生态成熟

Qwen、DeepSeek等模型全面适配昇腾平台

4

工程化落地

企业级推理引擎

在CANN架构上规模部署

CANN：异构计算架构优势

02

CANN架构

通过分层设计理念,CANN提供了类似CUDA的完整开发体验,同时针对昇腾架构进行深度优化。

CANN

生态兼容性：无缝对接主流框架

框架适配能力

CANN向上兼容主流深度学习框架,开发者无需大规模改写代码即可将模型迁移至昇腾平台。

- **PyTorch** - 通过插件实现算子映射
- **TensorFlow** - 原生支持计算图编译
- **MindSpore** - 华为自研框架深度集成
- **ONNX** - 标准模型格式直接加载

CANN

核心优势：全链路性能优化

1

算子级优化

内置1400+深度优化算子,覆盖主流神经网络层,单算子性能接近理论峰值。自动调优引擎根据硬件特性动态选择最优实现。

2

图编译加速

Graph Engine执行全局优化,包括算子融合、内存复用、数据排布调整等,减少内存拷贝和kernel启动开销。

3

多卡并行

HCCL通信库提供高效集合通信原语,支持数据并行、模型并行等多种分布式训练推理模式。

CANN

技术选型：为何选择CANN与昇腾

政策合规

实现算力自主可控,规避供应链风险

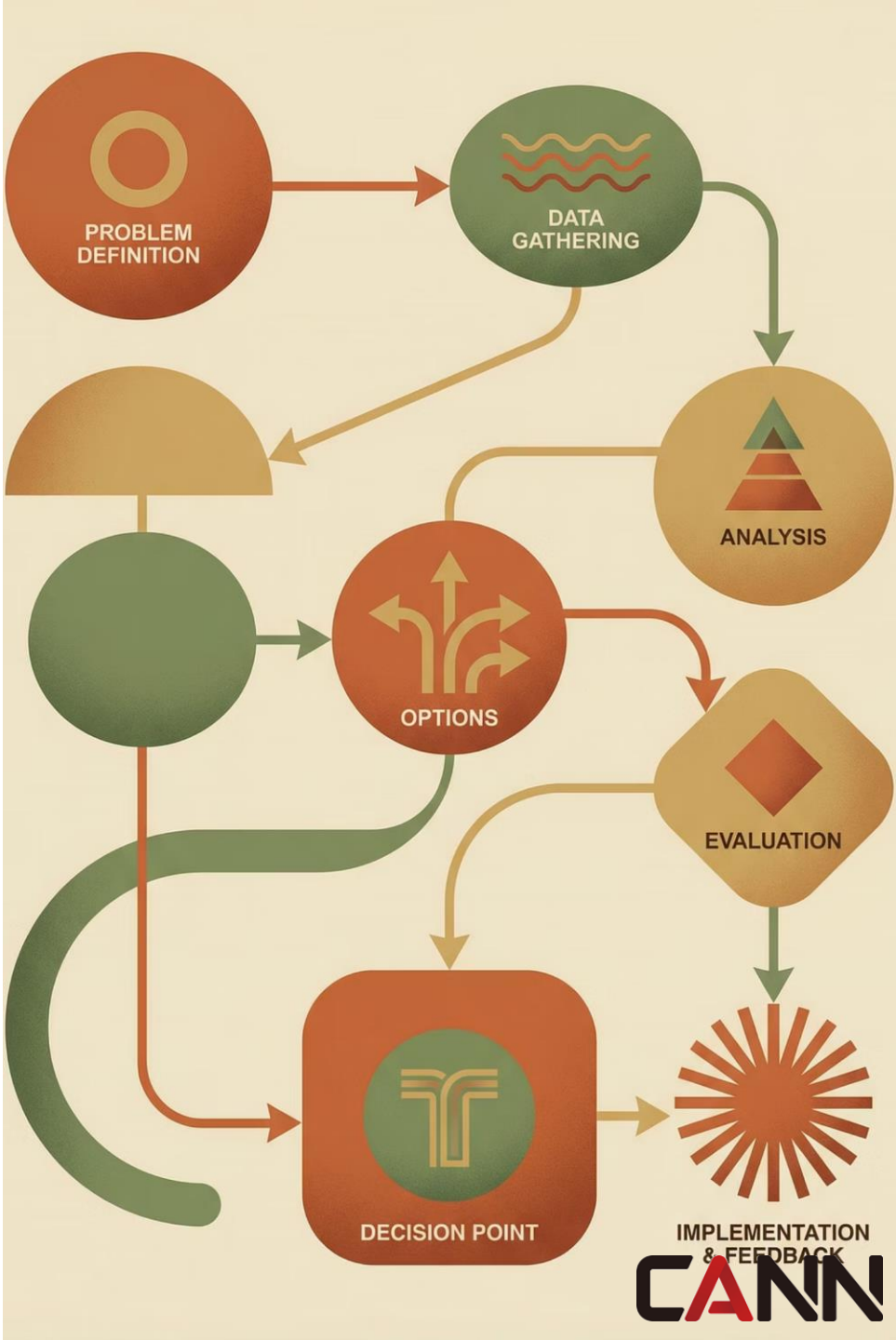
生态成熟

主流框架与模型全面适配,开发迁移成本可控

性能保障

深度软硬件协同优化,推理效率满足生产需求

> 综合政策导向、市场格局与技术成熟度,CANN架构与昇腾NPU成为我们招投标领域大模型推理的最优选择,为后续长文本优化实践奠定坚实基础。



招投标场景的大模型推理困境

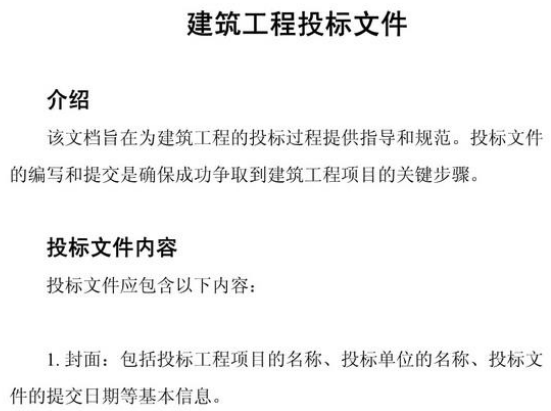
03

招投标文件的复杂性挑战

业务场景特点

招投标文件往往**篇幅极长**,包含各种格式的复杂内容。这些文件不仅有大量正文文本,还可能包含**表格、图像(如盖章的扫描件、技术图纸)、公式符号**等多模态信息。

这对AI模型的理解和处理能力提出了极高要求:模型需要具备**超长文本理解**能力,以及强大的**多模态处理**能力,才能完整地解析招投标文件中的关键信息。



建筑工程投标文件

介绍

该文档旨在为建筑工程的投标过程提供指导和规范。投标文件的编写和提交是确保成功争取到建筑工程项目的关键步骤。

投标文件内容

投标文件应包含以下内容:

1. 封面:包括投标工程项目的名称、投标单位的名称、投标文件的提交日期等基本信息。

文本量

单个文件可达数万至百万token级别

多模态

混合文本、表格、图像、公式等多种格式

CANN

GPU到昇腾NPU的迁移挑战

01

环境与兼容性问题

某些在GPU上使用的模型算子在Ascend平台上最初可能缺乏直接支持,需要开发者重新实现或替换算子,显著增加了开发成本和时间投入。

02

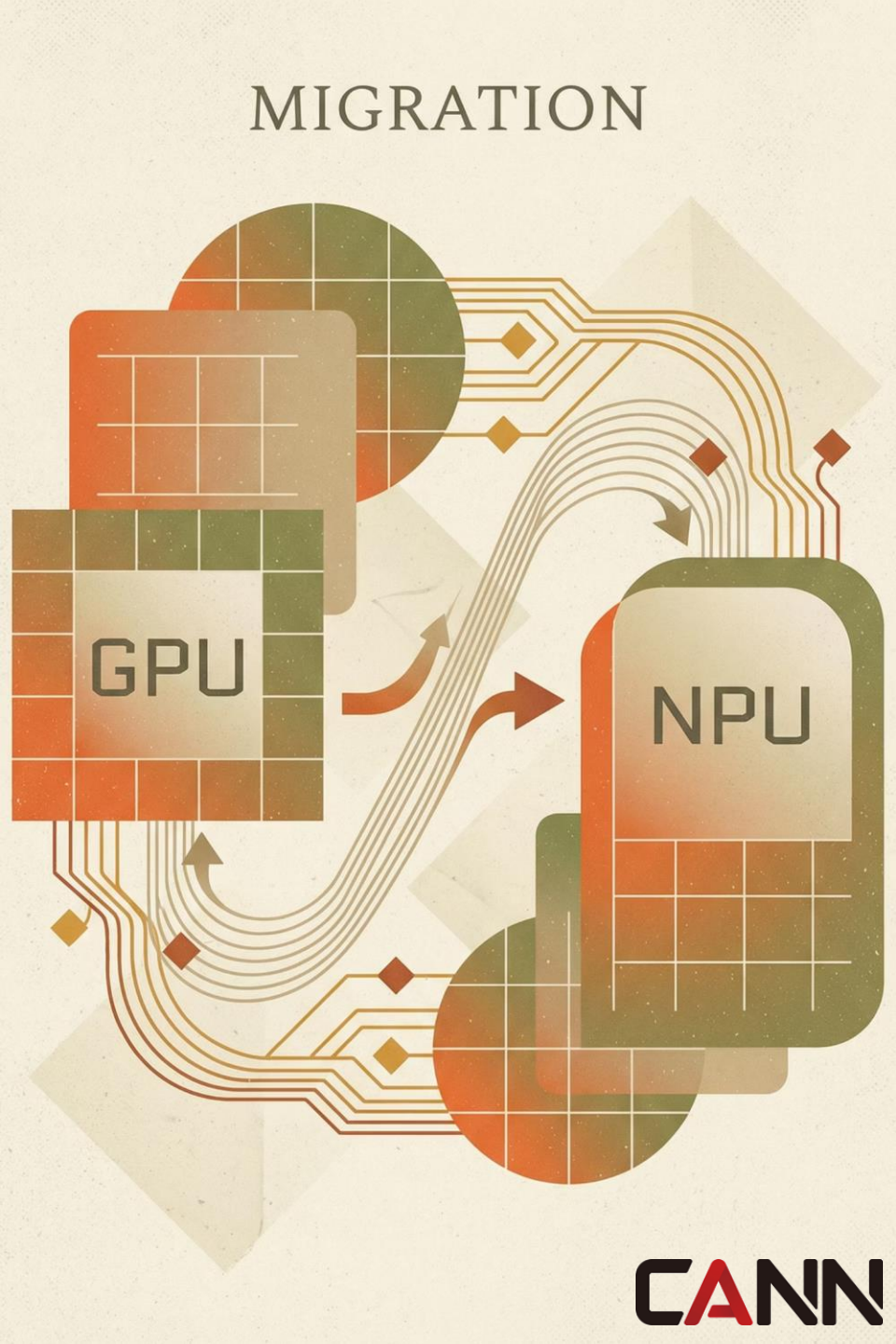
软件栈差异

Ascend软件栈(MindSpore或PyTorch+torch_npu/ONNX Runtime)与原有CUDA栈存在明显差异,开发团队需要较长的学习曲线才能上手。

03

性能优化不足

由于软件优化不到位,初期跑大模型时无法充分利用NPU算力,导致性能未达预期,出现"有力使不出"的困境。

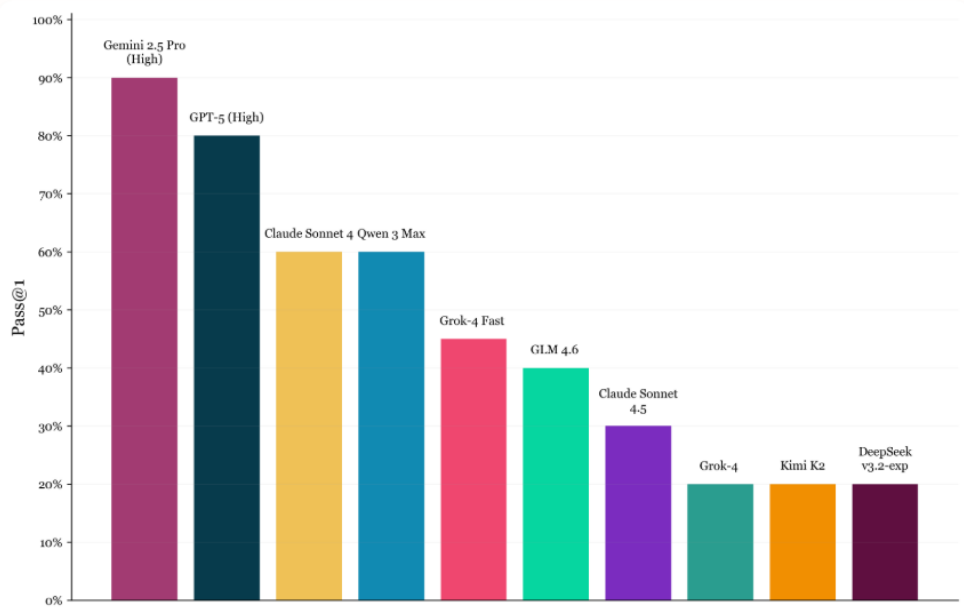


超长文本处理的技术瓶颈

计算复杂度爆炸

传统Transformer的自注意力机制计算开销为O(n^2),当输入长度n非常大时(例如上万甚至百万token),显存和计算量都难以支撑。

What do 1M and 500K context windows have in common? They are both actually 64K.



Pass@1 scores on the 128k subset of LongCodeEdit.

Pass@1 在 LongCodeEdit 的 128k 子集上得分。

核心问题

显存耗尽

长序列处理导致显存占用激增,超出硬件承载能力

运算速度骤降

推理延迟严重,甚至无法加载超长序列

系统瓶颈

1M字节级超长文本推理成为业界痛点

多模态解析的技术难题

OCR与版面分析

需要先对图片执行OCR和版面分析,提取文字和结构信息

信息融合

将提取结果与文本一起输入模型,进行语义对齐和理解

推理输出

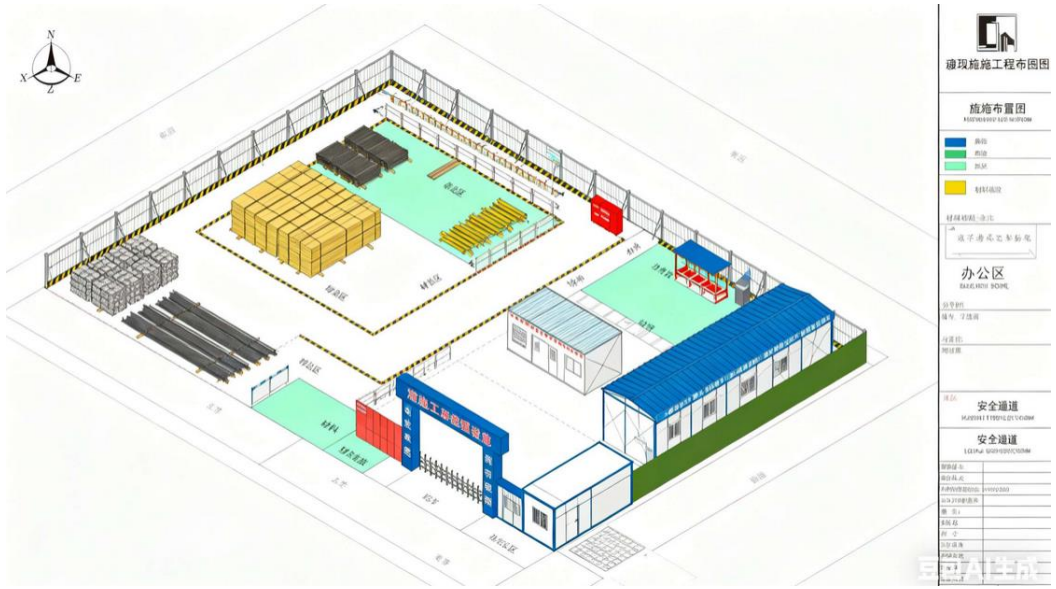
多阶段流程衔接不当会带来额外延迟和误差传播

纯文本大模型无法直接读懂扫描图片中的文字或表格结构,需要融合OCR、图像识别等能力。不同模态的信息在语义上如何对齐也是重大难点,模型需要将图文信息深度融合后才能进行正确的推理。

CANN

招投标场景的多模态信息类型

投标单位
投标报价(含税)(%)
投标保证金(元)
投标截止日前最新一期江西省发展改革委关于成品油价格调整的通告中柴油最高零售单价（元/吨）作为基准价(元)
投标单价 （元/吨）(元)
投标总报价 （含税）（元）(元)



表格数据

复杂的报价表、技术参数对比表、资质表格等结构化信息

印章与签名

扫描件中的公章、签名等关键认证信息需要识别验证

技术图纸

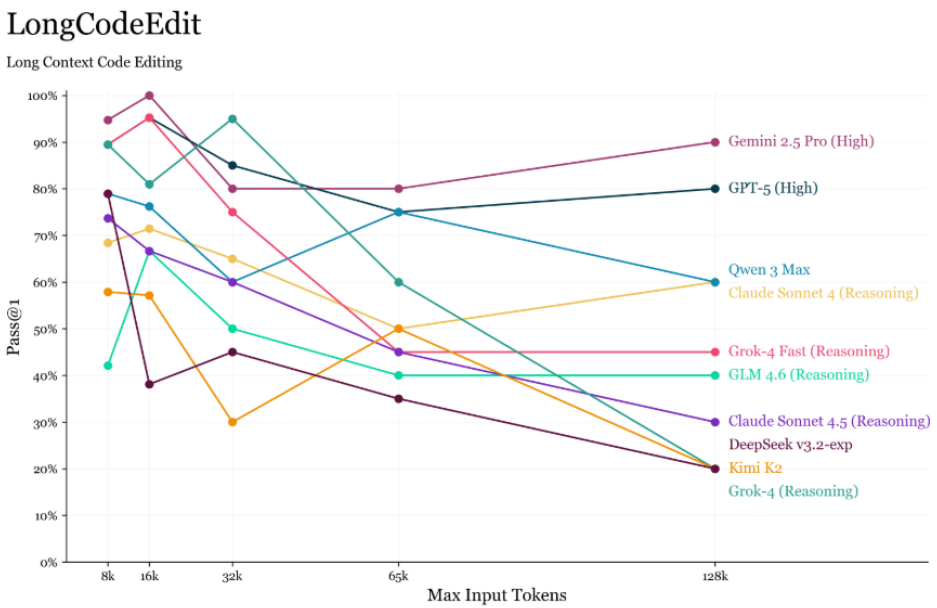
工程图纸、设计方案图、流程图等专业图像信息

公式符号

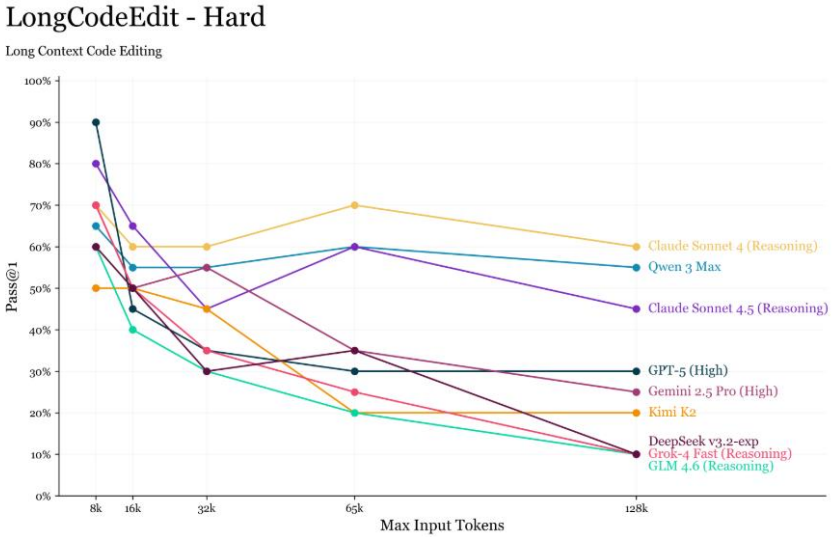
技术文档中的数学公式、化学式等专业符号表达

CANN

性能与成本的双重制约



LongCodeEdit Scores LongCodeEdit 评分



LongCodeEdit - Hard Scores

3-5x

推理耗时

超长文本处理相比普通文本耗时显著增加

10GB+

内存占用

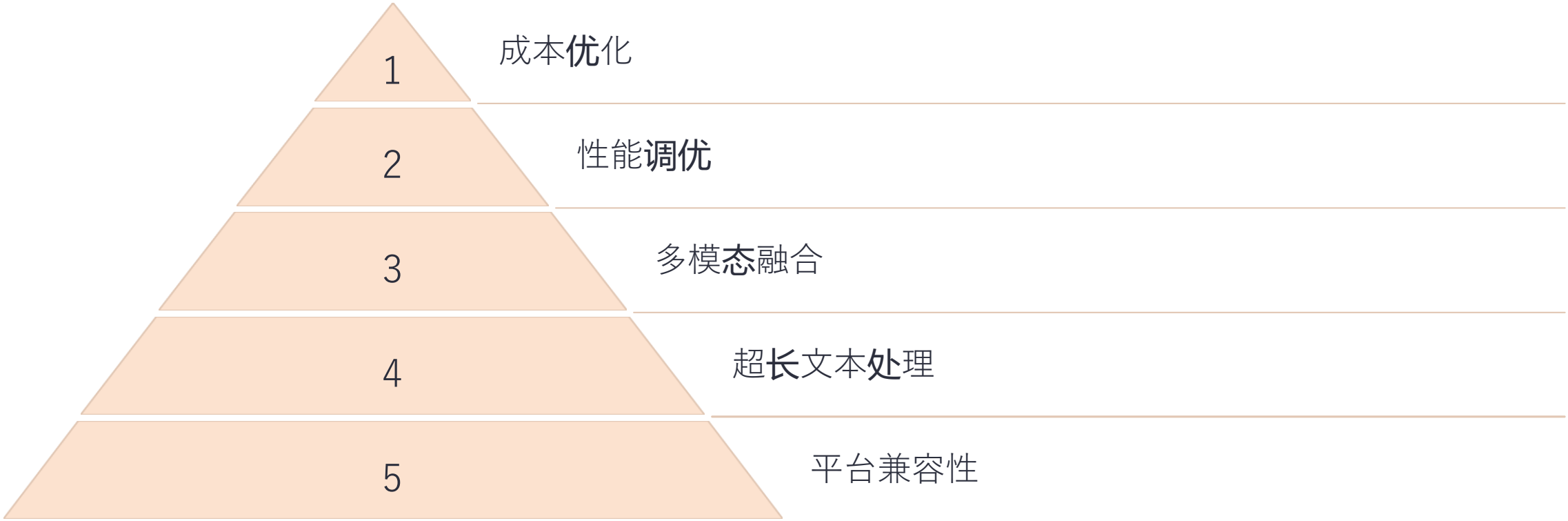
单次推理内存占用可能超过10GB

企业面临的困境

- **吞吐量低下:** 单个招投标文件解析耗时过长,难以满足批量处理需求
- **算力投入高:** 可能需要部署多卡NPU或大型集群才能勉强应对
- **人才缺口:** 缺少专门的优化人才来处理如此复杂的模型
- **成本压力:** 一般企业难以承受高昂的硬件和人力成本

CANN

技术挑战的层次结构



从底层的平台兼容性到顶层的成本优化,每一层都需要系统性的解决方案。基础问题不解决,上层优化就无从谈起,形成了层层递进的技术挑战体系。

核心技术难题总结

平台迁移

GPU到昇腾NPU的环境兼容性与算子适配

长文本

百万token级别的显存管理与计算效率

多模态

图文表格的统一理解与语义对齐

性能成本

推理效率与硬件投入的平衡优化

CANN

应对思路：通用大模型优化推理作为企业首选方案

04

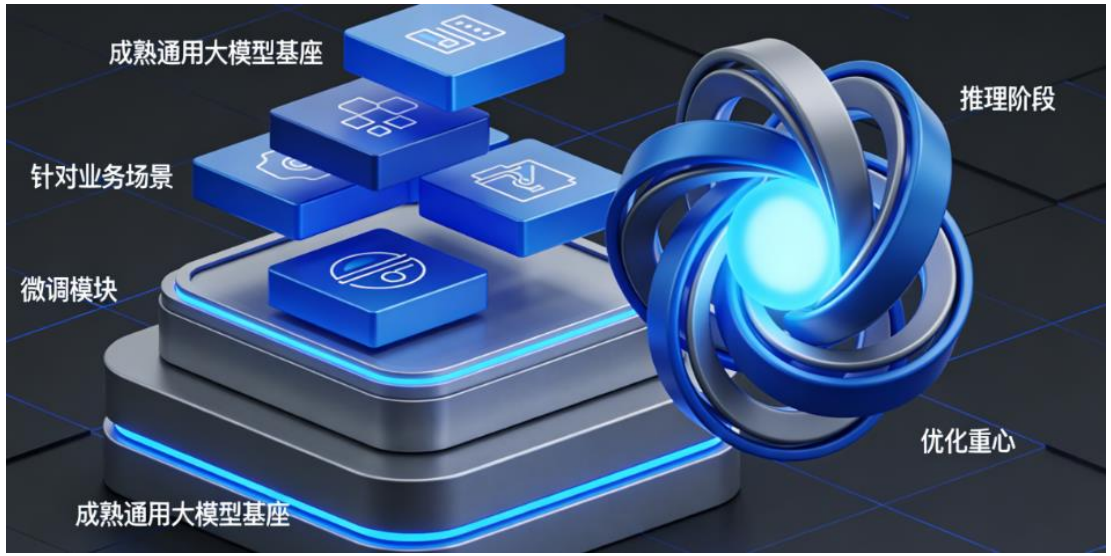
核心策略：以小搏大,优化先行

面对大模型落地的成本与算力挑战,我们不选择从零构建"大而全"的模型体系。

核心价值主张

- **跟上技术发展**:站在巨人肩膀上快速迭代
- **符合企业实际**:匹配资金和人力约束条件
- **性能与成本平衡**:达到实用效果同时控制投入

> **通过一系列推理优化手段,在CANN架构上实现大模型应用的最佳实践**



核心思路是:**选取成熟的通用大模型作为基座,针对业务场景进行适度微调,然后将优化重心放在推理阶段。**

CANN

方案优势一:显著降低研发成本

复用成熟基座模型

利用开源模型(LLaMA、ChatGLM)或国内开放模型(DeepSeek、星火等)作为起点,企业无需巨资从头训练

降低人才与算力门槛

只需在推理端进行优化和少量微调,对算法专家和大规模算力集群的依赖大幅减少

快速迭代验证

基于成熟模型架构,企业可以更快完成业务场景的适配和效果验证,加速上线周期

方案优势二:大幅降低算力开销

优化推理的核心价值

- 使用较小规模模型达到接近大型模型的效果
- 支持低精度计算,减少硬件资源需求
- 在保证性能的前提下优化推理速度

以小搏大成为可能:经过优化的通用模型在垂直领域表现已可超越GPT

大模型压缩实践

130B	<3%	128K
压缩后参数规模	效果损失	长文本处理能力
通过剪枝和蒸馏技术	在压缩后仍保持高精度	支持超长上下文解析

通过模型压缩将参数规模从更大模型缩减至130亿,在效果几乎无损的前提下,大幅提升了长文本解析效率。这证明中等规模模型配合优化完全可以胜任专业任务。

在招投标等专业场景中,选择经过优化的通用大模型,**可用更少的Ascend芯片完成推理任务**,显著减少设备投入。

CANN

方案优势三:弹性利用云端算力

推理即服务

通过华为云ModelArts等平台,按需调用昇腾集群进行大模型推理

按需分配算力

无需自建庞大硬件基础设施,根据业务负载弹性扩展计算资源

成本优化

算力成本按实际使用量付费,避免固定资产投入和闲置浪费

CANN

云服务模式的核心价值

降低准入门槛

企业无需深厚的AI基础设施建设经验,即可快速获得大模型推理能力

CANN环境支持

云端提供完整的昇腾AI计算架构,开箱即用的优化环境

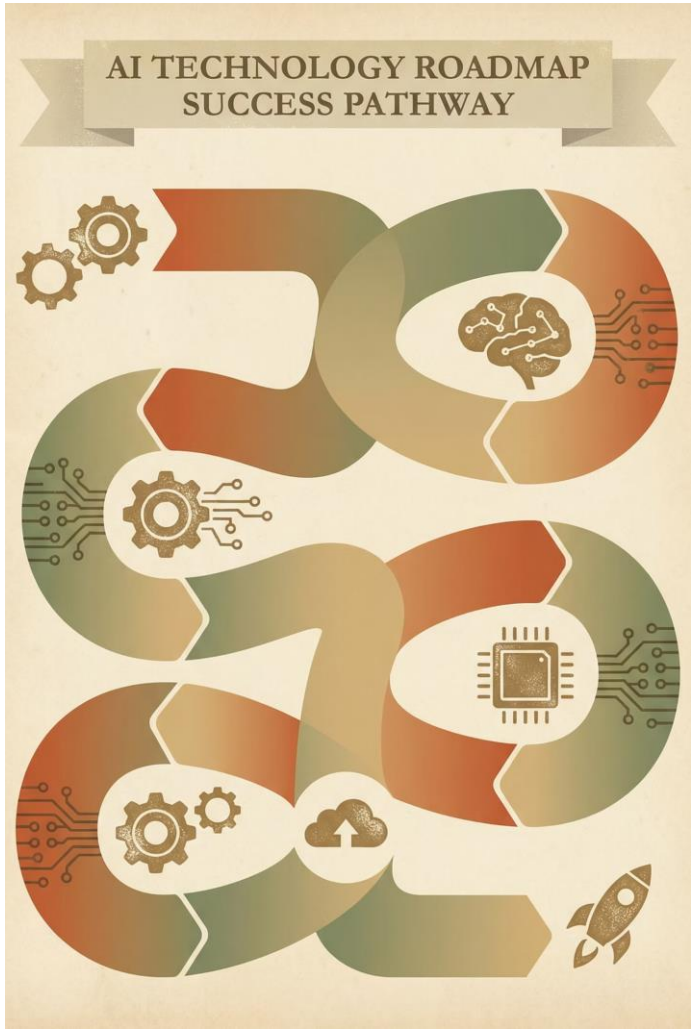
应对峰值需求

在业务高峰期自动扩容,平滑应对超大规模或突发负载

这种云服务模式让中小企业也能享受到大模型技术的红利,真正实现**算力资源的民主化**。

CANN

完整技术路线总结



01

选择预训练大模型

开源或商业通用/行业大模型作为基座

02

场景化微调

针对招投标等业务场景进行适度调优

03

推理端优化

模型压缩、量化、加速等技术提升推理性能

04

云端弹性部署

利用CANN架构和云服务实现灵活扩展

CANN

长文本推理的优化方法实践

05

长文本推理优化的三大维度

模型架构与算法优化

1. 长上下文模型与分段策略演进
2. 稀疏注意力机制
3. 动态长度支持等核心技术创新

CANN架构层面优化

1. CANN图编译优化与算子融合
2. 精度与数据类型优化策略
3. 多卡并行等底层加速

多模态融合优化

OCR加速、版面分析与图文融合处理的端到端解决方案

CANN

长上下文模型与分段策略演进

第一版方案:分段处理+汇总

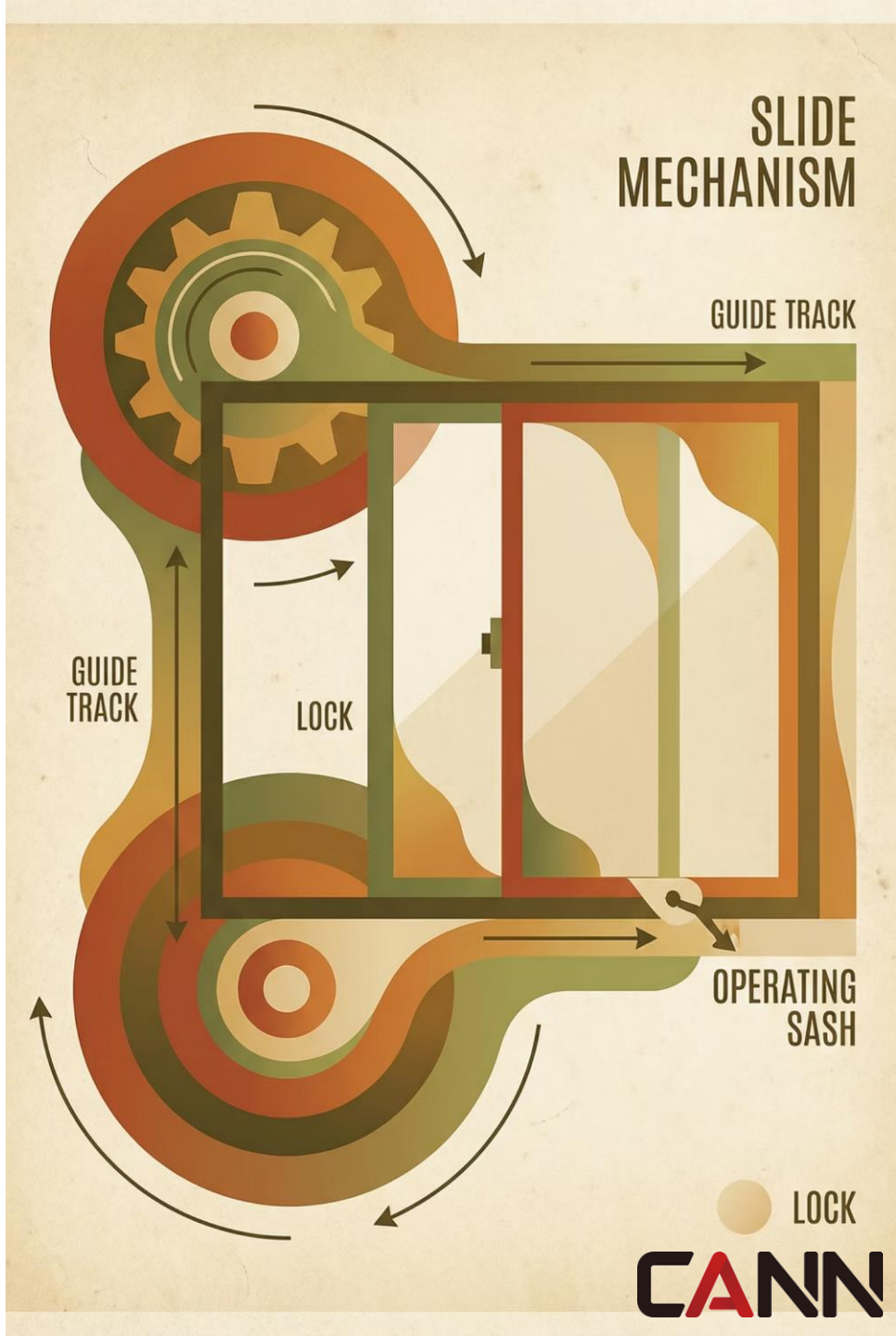
- 将长文档按章节或段落划分
- 分别提取要点或生成摘要
- 最后整合中间结果
- 避免单次推理内存爆炸

局限性: 对颗粒度文本检测密度和准确性要求高的场景效果不足

第二版方案:重叠滑窗+RAG

- 重叠滑窗机制保留相邻段落重复内容
- 确保模型了解跨段信息
- 结合检索增强生成(RAG)思路
- 多段重复对比提升准确性

突破: 显著提升了文本检测的精准度和完整性



稀疏注意力机制的突破性优化

01

问题识别

长文本Transformer中自注意力模块是主要的计算瓶颈,计算复杂度呈二次方增长

02

解决方案

引入稀疏注意力机制,有选择地关注关键位置而非全量计算注意力权重

03

Top-K分块注意力

智能识别对当前任务最重要的Top-K内容块,重点计算这些块之间的注意力

04

性能突破

单卡顺畅处理超过百万Token的长文本,完全突破长序列处理的显存与性能瓶颈

实践效果: 对长文本的注意力计算进行裁剪和近似后,加速比非常显著,同时保证整体效果

CANN

动态Shape技术实现灵活推理

传统模型按最大序列长度进行Padding,造成短文本场景下的算力浪费。动态Shape技术彻底解决了这一痛点。

开启动态Shape
ATC工具编译时启用动态shape选项

接受可变输入
模型接受长度可变的输入,无需固定最大长度

提升吞吐
短文本按实际长度运行,避免处理无效填充

动态批处理
根据并发请求数灵活调整批大小

验证结果: 小批量短文本请求快速返回,大批量长文本高效并行处理,达到高NPU利用率

CANN

CANN图编译优化与算子融合

图级优化

逐元素操作融合到矩阵运算中,减少中间内存读写开销

30%+

算力利用率提升

Ascend芯片算力利用率提升超过30%

多头注意力融合

将多头注意力的诸算子融合为整体内核,显著降低调用开销

2x

节点数量减少

经图优化后计算图节点显著减少

LayerNorm融合

LayerNorm、Dropout等算子与主计算融合,降低内存占用

40%

显存峰值降低

长文本场景下显存占用明显下降

CANN

精度与数据类型优化策略

1

FP16/BF16半精度

权重和激活转换为半精度,计算吞吐提升一倍,显存占用减半

2

W4A8量化

权重INT4、激活INT8,模型体积压缩至FP16的25%,极大缓解显存压力

3

混合精度拆分(MSD)

INT8激活拆解为高低位INT4分量,避免中间还原,进一步降低KV缓存开销

量化效果验证

- 长文本摘要任务吞吐提升约2倍
- 准确率下降不到1%
- 单卡可支持的Batch Size和序列长度上限显著提升
- 在有限硬件资源下部署大模型成为可能

多卡并行与流水线架构

模型并行方案

将模型不同层分布到多块NPU上,每块负责一部分计算。例如两张Ascend卡各算Transformer的一半层数,可容纳双倍规模的模型或上下文长度。通过HCCL高效通信库减少同步开销。

流水线并行/分阶段推理

将推理过程拆分为Prefill和Decode两个子系统并行运行。第一张卡负责长文档编码和预处理,第二张卡负责生成解码,两者以共享KV缓存连接。利用昇腾芯片高速全互联通信总线实现缓存高效共享。

核心优势: 避免单卡独自承担全部内存开销,充分并行长文本处理各阶段,达到缩短单次推理延迟的目的

CANN

多模态融合的端到端优化

OCR及图像预处理加速

- 使用昇腾NPU优化的视觉模型
- DVPP数据预处理加速库处理图像解码和resize
- 版面分析部署在Ascend上,借助ops-cv算子库加速
- 实现流式并行:NPU一边解析图像一边送入后续推理

模块化融合策略

- 表格数据以Markdown表格形式嵌入
- 印章图像以特殊标记替代
- 结构化解析输出JSON格式数据
- 微调模型学习处理多模态标记

性能提升: OCR在NPU上执行速度相比CPU有数量级提升,整个图像到文本提取过程实现了高效流式并行处理

CANN

实验效果与收益

06

优化成果概览

通过系统性优化，我们在昇腾CANN平台上实现了招投标长文本解析的全面性能突破

100万+	39%	2-3倍	50%+
Token处理能力	吞吐量提升	整体性能	时延降低
单卡支持超百万Token超长文档推理	推理吞吐量显著提升	相比业界主流的吞吐提升	单轮问答响应时间缩短

推理时延与吞吐优化

响应时间

在典型招投标问答任务上（几十页文档，提取关键信息问答），优化后模型的单轮问答响应时间降低了约50%以上

技术应用

动态batch和流水线并行等技术的应用，让NPU资源利用更加充分，实现最优性能

批量处理

整体吞吐量达到原GPU部署的2-3倍，支持业务在有限时间内完成更多任务

内存占用与成本优化

精细内存管理

通过量化策略和内存优化，峰值显存占用控制在较低水平

显存占用下降

FP16精度模型量化为INT8后显存占用下降约60%，可在Ascend 310等小显存推理卡上运行

硬件资源节省

很多场景已由原来的双卡推理缩减为单卡即可完成，节省硬件投入

成本优势明显

每处理一份招标文件所耗的算力资源成本下降约40%

业界对比与验证

参考业内领先实践，佐证优化方案的实际价值

星火长文本大模型

通过模型剪枝和蒸馏，将模型压缩到原有规模的一半以下，处理10K到128K长文本时性能居业界最优，首token响应时间大幅缩短

昇腾社区MoE推理优化

借助Grouped Matmul和混合精度INT4/8方案，使得MoE大模型的Decode阶段性能提升了超过30%

这些数据与我们的实验结论一脉相承：通过软硬协同的优化，大模型长文本推理完全可以实现又快又好的效果

CANN

优化实践的核心价值

01

性能突破

克服GPU时代令人生畏的长文本、多模态推理难题，单机NPU即可流畅处理复杂招投标文件

02

成本控制

算力资源成本下降约40%，硬件投入显著减少，为企业带来实实在在的经济效益

03

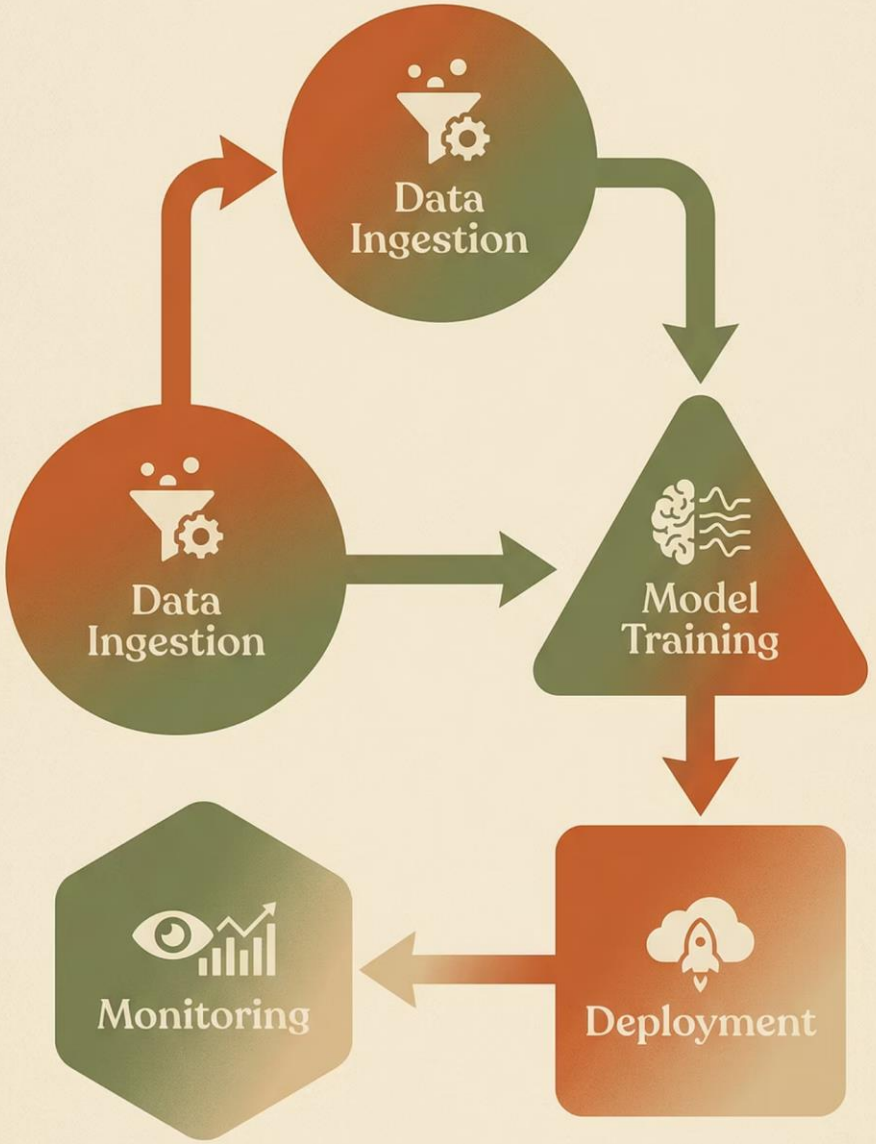
软硬协同

充分证明了软硬件协同优化的价值，借助CANN完备的软件生态实现技术突破

04

AI落地

让普通企业也能驾驭大模型，推动AI真正落地业务场景



CANN

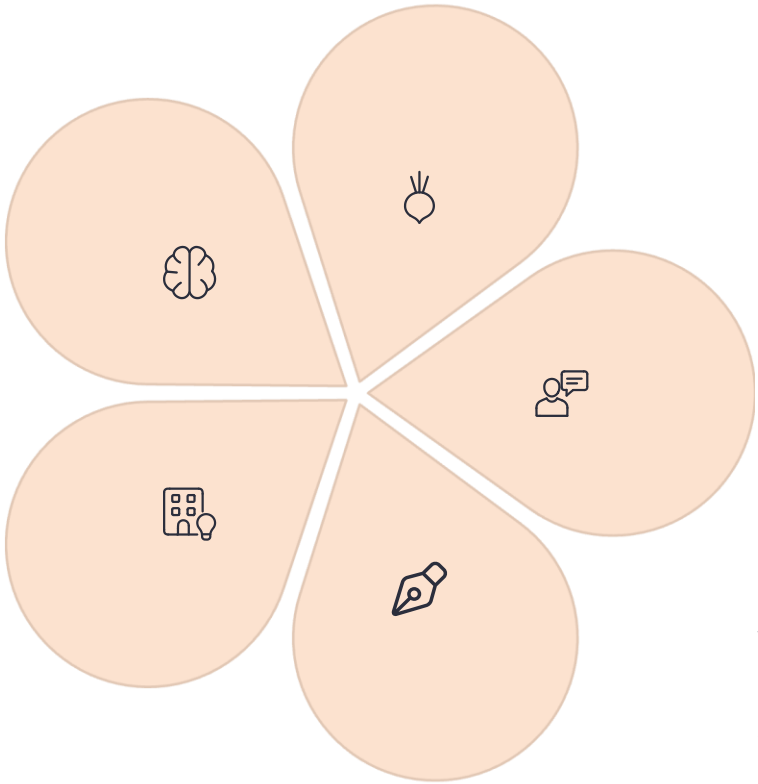
未来技术展望

智能图优化

更加智能的图优化器与高效算子实现

架构创新

适配昇腾架构的新型模型结构



多模态大模型

端到端文档智能解析能力

超长上下文

社区最新超长上下文处理技术

低比特量化

新算法助力性能持续优化

CANN

Thanks！

访问CANN开源社区

关注昇腾CANN公众号