

盘古大模型：AI4Industry

- ▶ 华为公司一直高度重视AI的作用，任总早在2012年就亲自指示成立了诺亚方舟实验室，专注于从事AI研究。
- ▶ 华为提供包括芯片设计、底层开发、高层应用框架、以及最终产品的全栈技术，包括昇腾910AI芯片、Atlas900 AI集群、昇思MindSpore AI开发框架、ModelArts AI全流程开发平台等等，为客户提供最全面和最彻底的解决方案。
- ▶ 2024年，华为云发布基于全栈自研技术开发的盘古系列大模型盘古5.0，构建三个层次的系列大模型：

L2 场景大模型	政务热线 慧眼识事	网点助手 财务异常分析	供应链物流 器件分配	先导药物筛选 小分子优化	传送带异物检测 掘进序列检测	铁路TFDS检测	台风路径检测 海浪预测
L1 行业大模型	政务 大模型	金融 大模型	制造 大模型	药物分子 大模型	矿山 大模型	铁路 大模型	气象 大模型
基础大模型							
盘古语言大模型				盘古多模态大模型			
1.5B 7B 38B 135B				多模态理解 图像生成 视频生成			

如何在昇腾集群上训练出千亿大模型？

高效模型训练

高效增量训练

内存节省4倍+, 训练加速30%

多维混合并行

实现超大规模AI集群近线性加速比

3D序列并行

吞吐提升7倍, 显存降低50%~80%

Semi-online RLHF

迭代效率提升4倍

部署优化降本增效

QuantGPT量化技术

内存降低2-4倍, 推理加速15-30%

投机式推理

推理加速50%

分离部署

推理集群吞吐量提升7倍

数据科学

数据工具链

AI处理工具链

先导实验

验证数据版本和合成数据准入

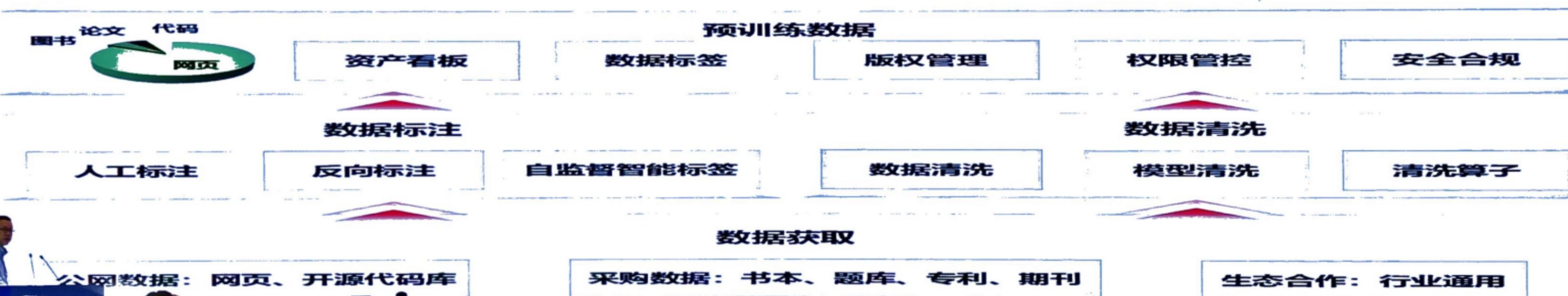
数据合成

提升长序列能力20%

数据课程

5层数据体系, 提升训练效率20%

预训练数据管理平台



数据合成：自动化构建高质量数据，提升复杂推理和长序列能力

SELF：自我评估+自我修正合成复杂推理数据

浅层反馈：评判

打分评价：局限性

- 很难从多个维度反馈回复质量。
- 没有学习一个时候选集中答案(A/B/C/D)更好的回复。

A labeler ranks the outputs from best to worst.
①-②-③-④



中层反馈：评判+解释

语言评价

- 不仅评判模型生成的内容好或者不好，还要解释为什么。

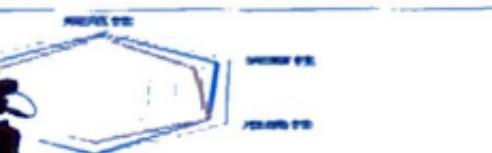
For response to explain neural networks to a child:
A is less preferred compared with B.
B is equally good as C, and
C is less informative than D.

深层反馈：评判+反馈+示例

给出好的示例

- 当模型给出不好的回复时，除了解释为什么好，同时还给出好的示例。

Please note that auto-tiny uses less tricks to elaborate but they can also be a lot of fun. But when it comes to challenging us to give bad product info or explain a consumer question for under \$1000 dollars, we just couldn't say no. Transforming a very smart system doesn't mean to lower your bridges. Convert things like this:



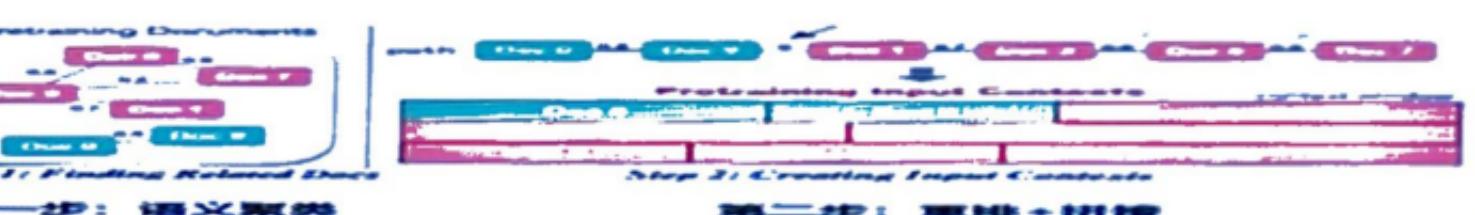
当前进展

合成长序列数据
大海捞针提升20%



长序列数据合成

- 长序列文本合成：聚类领域文档形成具备一定知识广度的基础数据，挖掘关联关系形成知识流转图，采用Multi-Hop QA等方法进一步构造深度问答数据，匹配基础文档构造**500B超长序列文本**，提升模型在“大海捞针”等任务上的思维能力



多维混合并行：实现超大规模AI集群近线性加速比

挑战

千亿大模型的训练内存占用多，计算量大，在训练中需要大量的梯度、参数等的同步，集群上无法实现线性加速比。

1024卡集群

计算和通信比例为**70:30**

2048卡集群

计算和通信比例为**40:60**

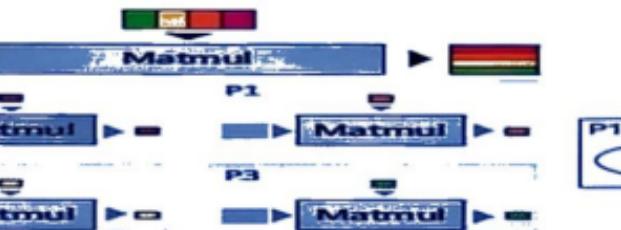


关键技术

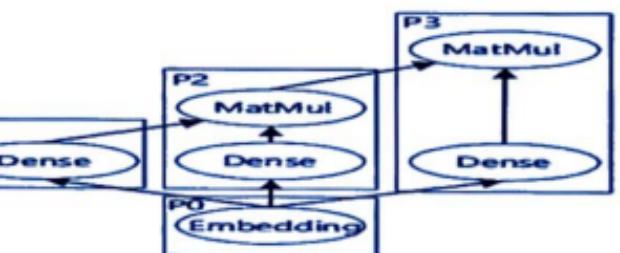
模型+集群的并行策略，模型切分更加平衡

- 计算-通信的高效流水线
- 局部重计算换取内存约简，降低模型切分压力

层内模型并行



层间模型并行



结果

实现集群**近线性加速比**

1024卡

90:10

2048卡

81:19

4096卡

76:24

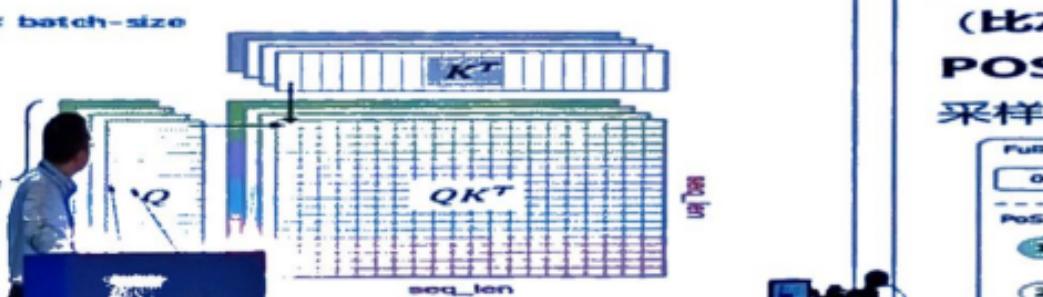
基于**通算融合**，**参数面**网络负载
均衡，**FFTS硬化调度技术**，
2023年底**4096卡**达到**90:10**

3D序列并行：提出多维混合并行算法，可以支撑38B模型128K长序列直训

挑战

长序列在训练和推理上会引入平方倍的复杂度。

盘古百亿-千亿模型**128K**序列训练，需要**5T以上显存**，传统并行策略显存溢出，并且计算速度慢，需要**4K序列8倍的时间**。



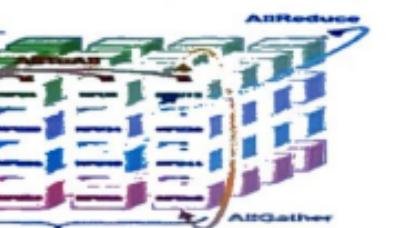
关键技术

- **3D序列并行：**提高通信效率和集群扩展性。
- **Pose 位置编码：**短序列模拟长序列训练。

3D序列并行

支持二维序列切分与一维模型切分，降低通信开销，并提高集群扩展性。

(比友商早1年提出)



POSE位置编码：通过在训练期间随机跳过一些随机采样的偏置项，可以短序列模拟长序列训练。



结果

支持**38B模型128K直训**，吞吐提升**1.4倍**，外推效果**接近全长微调**。

① 显存收益：显存接近线性下降

实验**512卡计算128k长序列**，对比**Megatron**吞吐提升**1.4倍**，静态显存减少**40%+**，动态显存减少**80%+**。

DP	SP	MP	静态显存 (MB)	动态显存 (MB)
1	1	8	31,828	51,904
1	2	8	26,032	26,272
1	4	8	21,968	13,312
1	8	8	18,168	6,840

② 吞吐收益

千卡吞吐从**10B/天**提升至**14B/天**，端到端吞吐率提升**140%**。

高效推理提升成本竞争力：内存降低1/4倍，推理加速50%，380亿已支持单卡推

挑战：生成模型参数量大，推理慢，占用内存高，端到端推理成本高。

显存占用和序列长度n成正比: 1750亿模型 4k长度

卷之三

非线性：生成模型直接使用非线性的量化算法会导致严重的精度下降

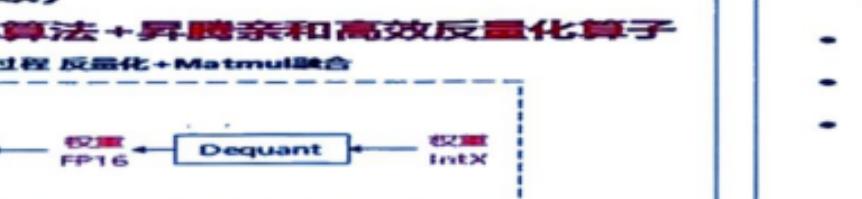
奇麗化（參照）

4/8-bit 权重量化算法

QuantGPT+昇騰垂和高效

反量化算子：模型内存降低

2-4倍，推理加速15-30%



2022 Outstanding Paper Awards

the 低比特圖像 (編存)

◎ cache 8-bit量化之压缩率最高达3倍。

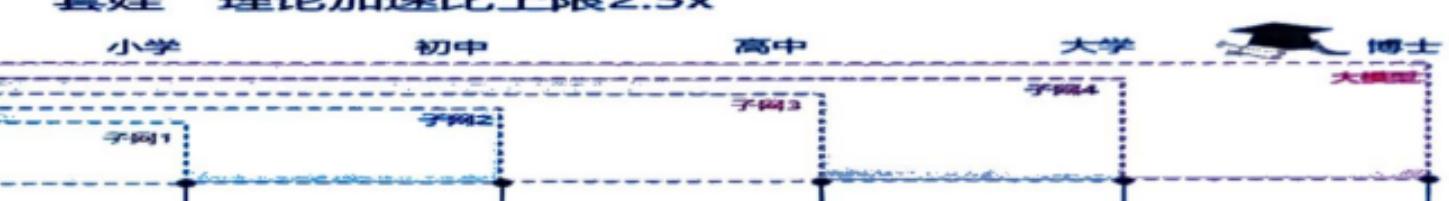


投机式

大小模型投射、小模型需要单独训练、维护、费时费力

“三”进归根机理

- 存占用少：只需部署一个模型
吐大：所有子网络共用前向推理
速比高：复用部分网络，预测匹配度高，叠加多级“套娃”，
“套娃”理论加速比上限 $2.5\times$



- 自投机小模型与大模型的token匹配率近
90%，高级投机式推理在盘古模型实现高达50%

FastAttention：通算融合&多样化算力&全序列并行，实现4M长序列推理

挑战

全量和增量推理时延和内存与输入序列长度呈**线性和平方复杂度**



- **内存瓶颈：**增量推理 $O(n)$ 内存复杂度，RingAttention序列并行方案多卡虽然降低了访存，但是增加了通信
- **计算瓶颈：**全量推理 $O(n^2)$ 计算复杂度
- **通信瓶颈：**全量推理在增量推理阶段，32B级别KV Cache通信

关键技术

FastAttention：高效Attention算子缓解长序列推理内存、计算、通信瓶颈，支撑长序列推理

FastAttention 1.0：通算融合加速FlashAttention2



FastAttention 2.0：多样化算力缓解内存瓶颈

- 全量推理：wave-front 并行，device 计算并下传 KVcache到host
- 增量推理：device+host 协同，CPU计算Attention

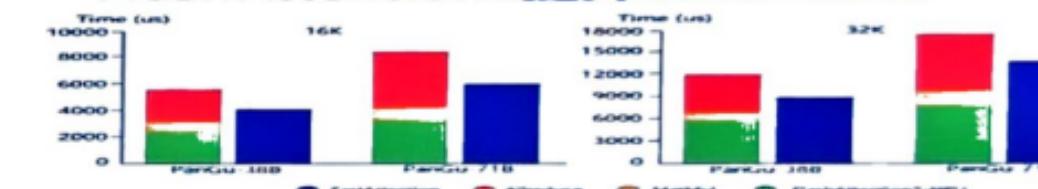
FastAttention 3.0：序列并行缓解通信瓶颈（开发中）

- 全量推理：3D序列并行切分KV Cache到多个NPU
- 增量推理：对Q进行通信，避免KV通信

当前进展

单机可推**38B的2M长序列**，多机推理**2.6B的4M序列**，未来扩展可以到**100M**。

- **FastAttention 1.0：** 推理速度相对 FlashAttention2 提升 **13%~28%**



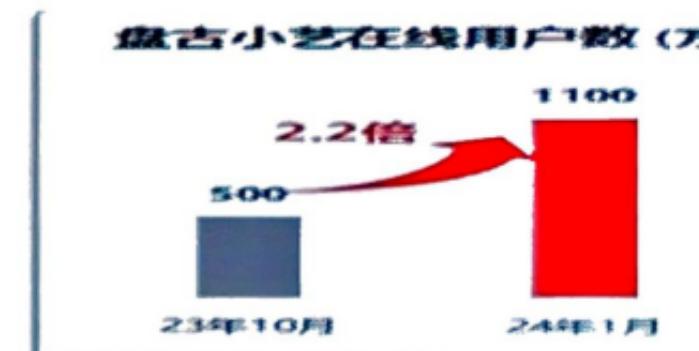
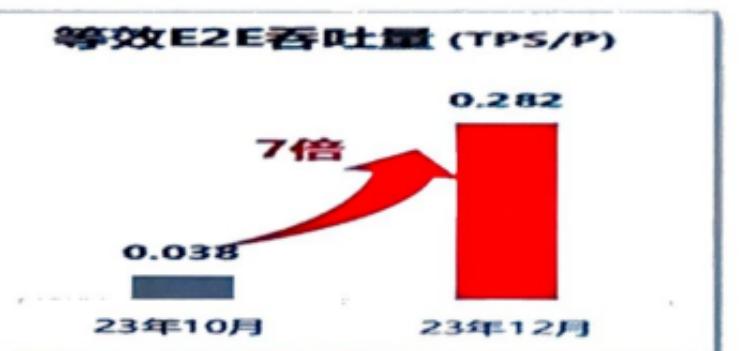
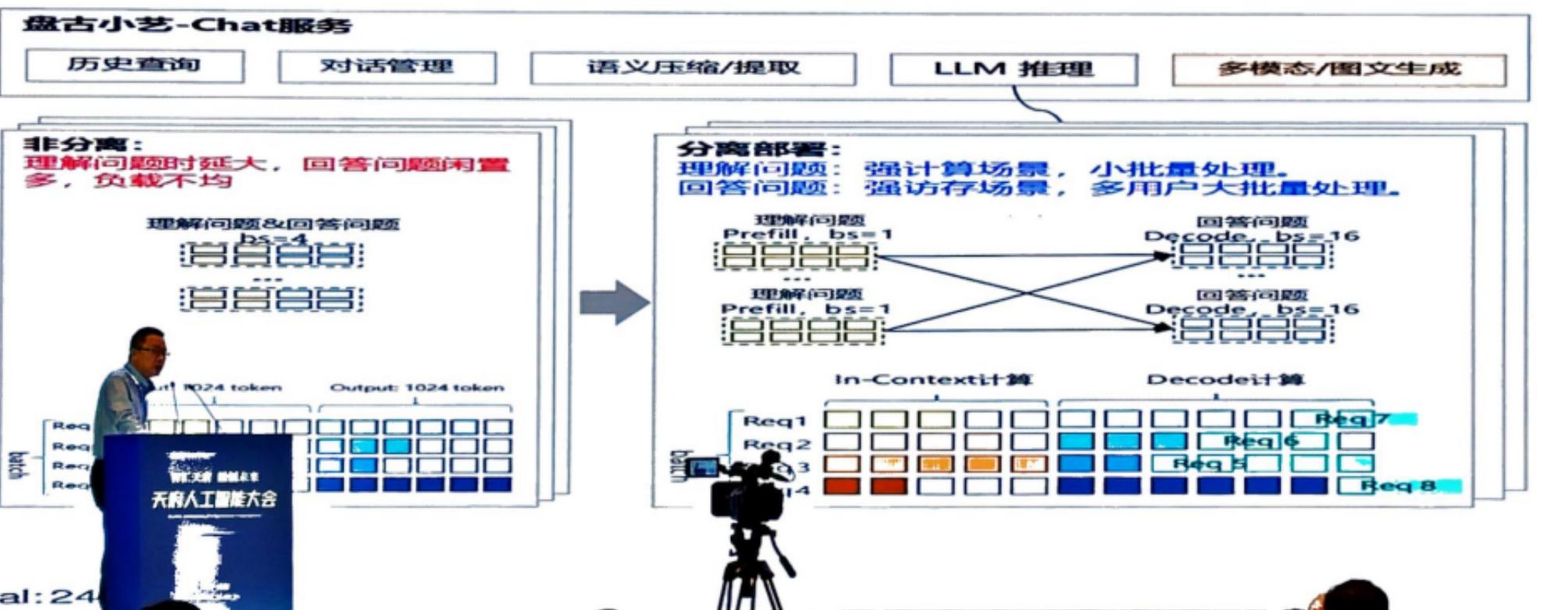
- **FastAttention 2.0：** host+pcie+device 协同，**单机8卡推理盘古38B 2M长序列**

- **FastAttention 3.0：** 通信量从 $BnLd$ 降为 $2BLd$ ，2.6B多机可推4M长序列

- **叠加其他压缩手段：** KV Cache量化压缩1倍；使用GQA进一步压缩5倍

分离部署：提升推理集群吞吐量7倍

在满足小艺业务体验时延约束下，AIGC/推荐场景通过集群分离部署，推理集群吞吐量提升7倍，支撑在线用户倍增



基于“分离部署”的持续优化

- 更灵活的模型量化：模型针对全量使用int4、增量使用int8的不同量化方案。预计提升模型吞吐10%以上。
- 支持灵活扩展分布式存储节点。支持多轮对话场景下的KV Cache复用，减低长序列下的时延。
- 对等架构，灵活配比：动态调整全量&增量的集群数目来应对业务忙闲时，以及后续的多模态场景。

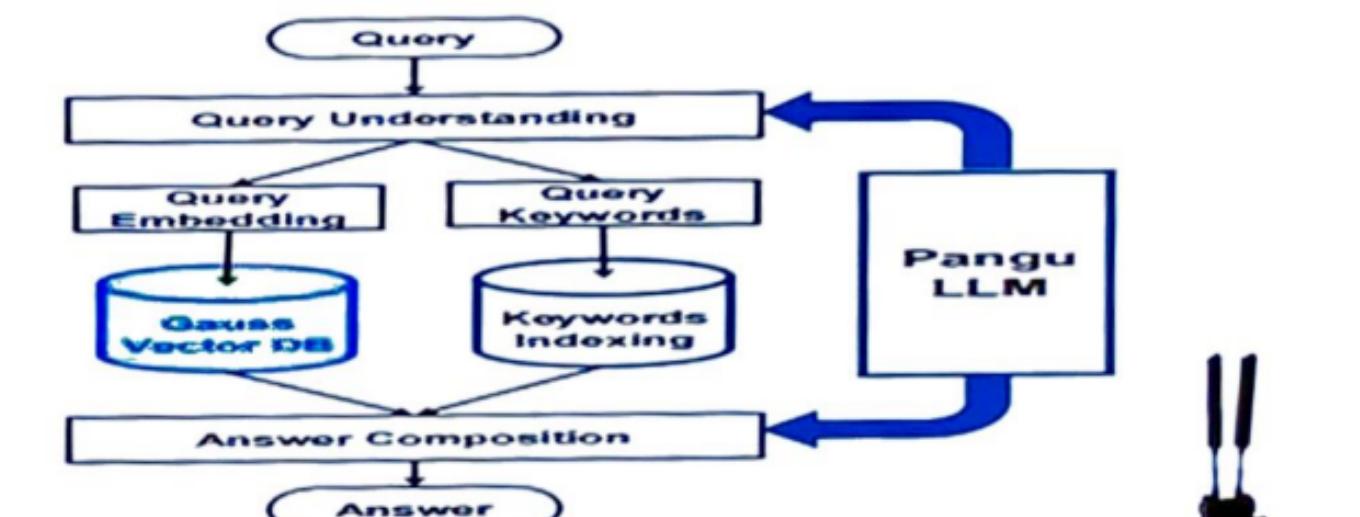
应用案例：大模型+向量数据库，电信领域知识问答达到专家水平

电信领域知识问答

- 数据规模巨大：4000万个文档
- 数据形式复杂，包含大量数字、表格、公式
- 问题复杂，需要精准知识问答
- 数字敏感

问题	HiSecEngine USG6603F-C天关产品中，光模块光纤的纤芯直径国际标准是什么？
标准答案	光纤的纤芯直径，有国际标准： 多模的为62.5um和50um，单模的为9um。
微调盘古模型 X	HiSecEngine USG6603F-C天关产品中，光模块光纤的纤芯直径国际标准是 125微米。
+向量数据库 ✓	HiSecEngine USG6603F-C天关产品中，光模块光纤的纤芯直径国际标准是 多模的为62.5um和50um,单模的为9um。

- 电信领域业务问答得分从**10.5分提升到96分**，达到专家水平。
- 五类工程师（包括通过数通HCIE 国际认证ICT专家），超过**95%专家**。



应用案例：大模型+天筹求解器，复杂问题建模和求解

Application Case: Supply Chain Optimization

- **Pangu Language Models:**
Problem Modeling

- Dialog
- Complex reasoning
- Code generation
- Summarization



- **OPTV AI Solver (Plug-ins):**
Problem Solving

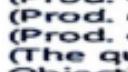
- Interpreting
- Calculation
- Optimization
- Math problem solving



For level-1 problems of general operation research, we reaches a comparable level with GPT-4.

Training data	32,108
Test data	289
GPT-4	0.7048
Pangu Production Model	0.6989

The supply chain planner needs to make decisions on the allocation of key components (raw materials), SC (structural component) 1 and 2. There are three requirement codes, namely INV (inverter) A, B, and C. The requirements are 50 pieces, 100 pieces, and 150 pieces, respectively. The virtual material cost is 1.3, 2.1 and 1.8 respectively. The inventory of raw material SC 1 and 2 is 100 pieces and 90 pieces, respectively. SC 1 can be used to produce INV A or B, and processing unit INV A or B requires two SC 1. Processing unit INV C requires one SC 2, and SC 2 can replace SC 1, but SC 1 cannot replace SC 2. May I ask how to make a decision on the production quantity of INVs A, B, and C, as well as the allocation of raw material SC 1 and 2, in order to maximize the virtual material cost of the requirements met? Build a planning model for this problem.



Variables:
(Prod. quant. of INV A) x
(Prod. quant. of INV B) y
(Prod. quant. of INV C) z
(The quant. of SC 2 replacing SC 1) r
Objective function:
(Maximizing total virtu. mater. cost) $1.3x + 2.1y + 1.8z$
Constraints:
(Max Prod. quant. constr. of INV A) $x \leq 50$
(max Prod. quant. constr. of INV B) $y \leq 100$
(max Prod. quant. constr. of INV C) $z \leq 150$
(consum. quant. of SC 1 cannot exceed invent.) $2x + 2y \leq 100$
(consum. quant. of SC 2 cannot exceed invent.) $z + r \leq 90$
(non neg. int. deci. var. constr.) $x, y, z, r \in \mathbb{N}$



The solution:
Solution status: OPTIMAL
Target value: 267.0
Decision variables:

Variable Name	Solution results
z	90.0
x	0.0
y	50.0
r	0.0

盘古大模型助力千行百业智能化：已落地10+行业的1000+AI项目

工业仿真：商飞“东方·翼风”

大客机机翼气动AI仿真
流场预测平均误差低至0.001星级
单次仿真速度提升1000倍
WAIC2023最高奖项SAIL奖

气象：登上《Nature》正刊

1.4秒完成24小时的全球气象预报
比传统方法快1万倍
误差降低12%以上
计算能耗降低60万倍
极端天气预报提升25%

工业制造：自动排产

产线分配计划从几个小时降至1分钟

EDA

大语言模型代码生成
测试样例生成覆盖率达到99.5%
端到端提升测试研发效率3倍+

药物

极大缩短药物研发周期
西交大附属医院1个月内发现了新的广谱抗生素

政务

自动调度后端上千应用系统
快速实现城市各类服务