



2024 AI+研发数字峰会

AI+ Development Digital summit

AI驱动研发迈进数智化时代

中国·上海 05/17-18

AI原生应用开发工具链详解

罗义云 阿里云人工智能平台

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



K+峰会

上海站

K+ 全球软件研发行业创新峰会

时间: 2024.06.21-22

K+峰会

敦煌站

K+ 思考周®研习社

时间: 2024.10.17-19

K+峰会

香港站

K+ 思考周®研习社

时间: 2024.11.10-12



K+峰会详情



AiDD峰会

上海站

Ai+研发数字峰会

时间: 2024.05.17-18

AiDD峰会

北京站

Ai+研发数字峰会

时间: 2024.08.16-17

AiDD峰会

深圳站

Ai+研发数字峰会

时间: 2024.11.08-09



AiDD峰会详情



罗义云

阿里云资深技术专家、PAI平台工程负责人

毕业于北京大学，曾任微软高级研发经理、旷视科技AI平台高级技术总监，现任阿里云资深技术专家、AI平台工程技术负责人。在机器学习、人工智能、大数据等方向有着深厚的技术积累和丰富的行业经验。

目录

CONTENTS

1. AI原生应用开发及其挑战
2. 成本和性能的考量：微调 vs. RAG
3. 效果优化的工具
4. 阿里云AI原生应用开发实践

PART 01

AI原生应用开发及其挑战

▶▶ 什么是AI原生应用

通义千问：AI原生应用（AI Native Application）是指那些从设计之初就深度融合人工智能技术，并将其作为核心功能和架构基础的应用程序。这类应用并非简单地在传统软件或服务中添加AI模块作为附加功能，而是从底层架构、开发流程、数据处理、交互方式乃至商业模式等多方面，全面、系统地运用人工智能来提升应用的智能化水平和服务效能。

▶▶ AI应用市场增长迅速

114%

AI应用下载量增加
(2023 H1 vs. 2022)

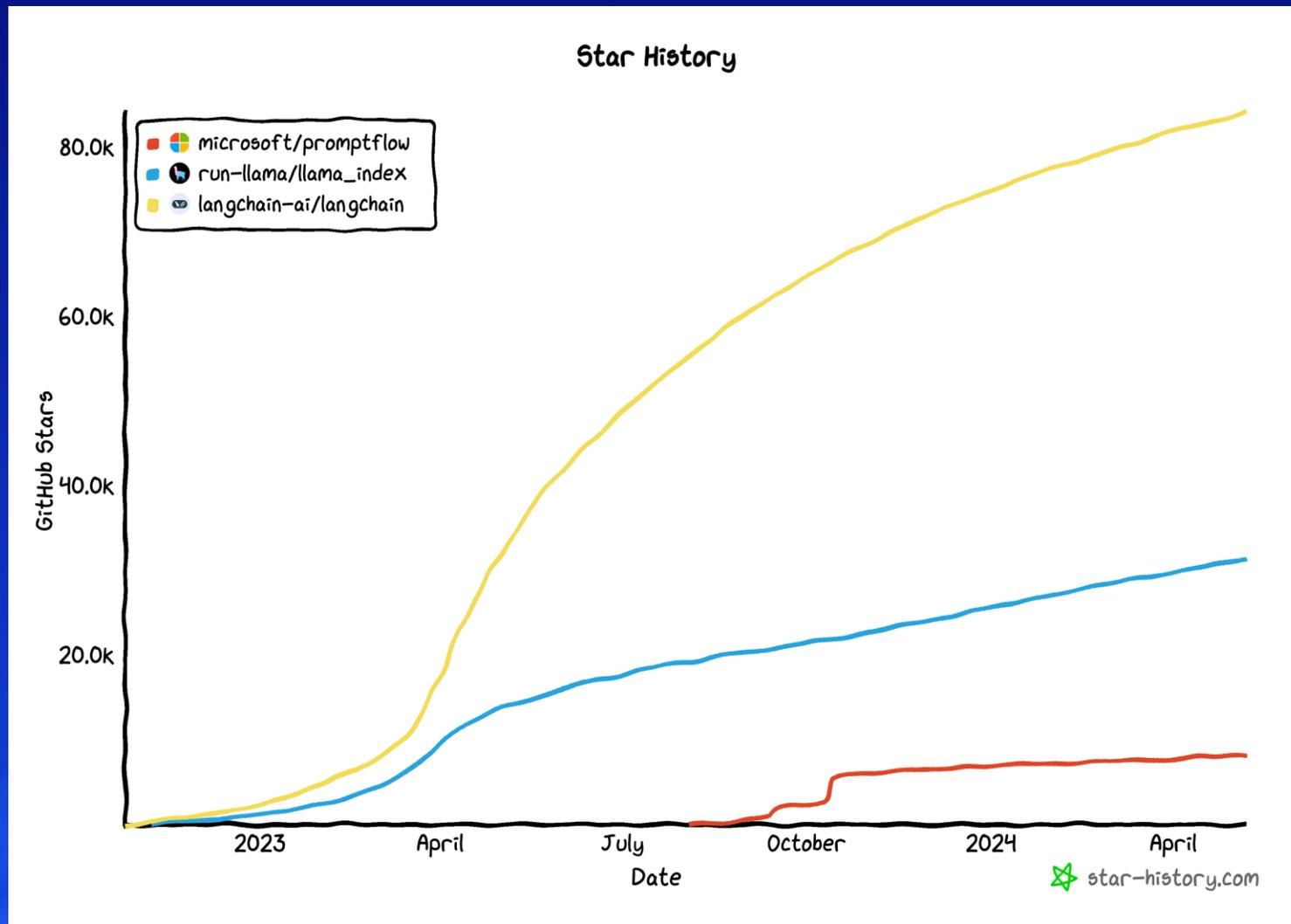
55.1%

2023 ~ 2027
生成式AI市场年复合增长率

5亿

2024
涌现的AI应用数量

▶ AI应用开发框架



▶▶ 典型的AI应用开发全流程



▶ AI原生应用开发的挑战

成本

- 训练成本
- 推理成本

效率

- 迭代次数多
- Debug困难

效果

- 幻觉
- 内容合规

PART 02

成本、性能和效果的综合考量： 微调和RAG

▶▶ LLM的应用面临什么挑战

- **推理成本/效率:**

- 大语言模型推理成本较高
- Prompt工程带来的响应延迟

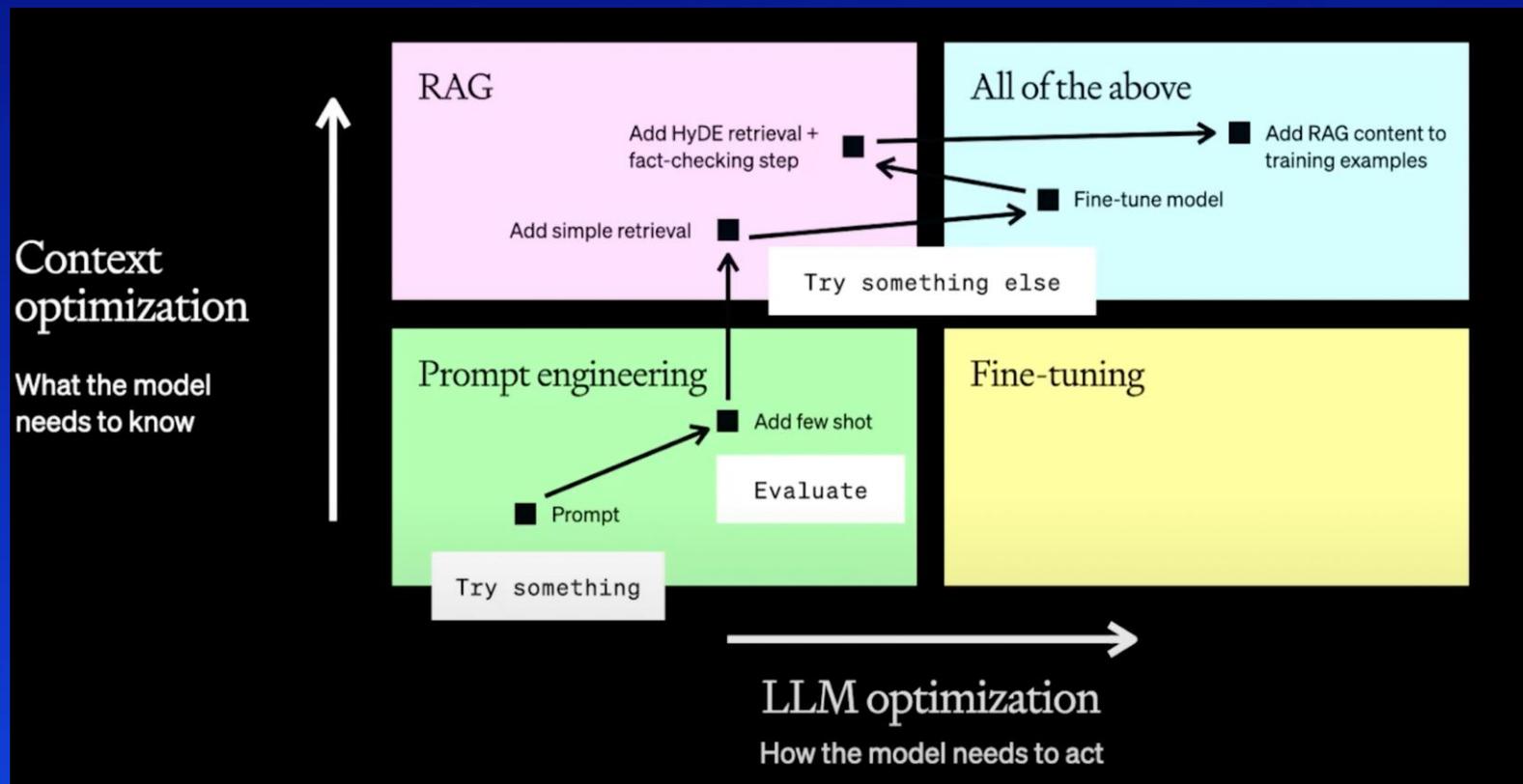
- **模型效果:**

- 缺少私有的长尾数据、实时的数据
- 模型存在幻觉
- 上下文长度有限

我们的目标: 增加知识、增加能力、减少成本

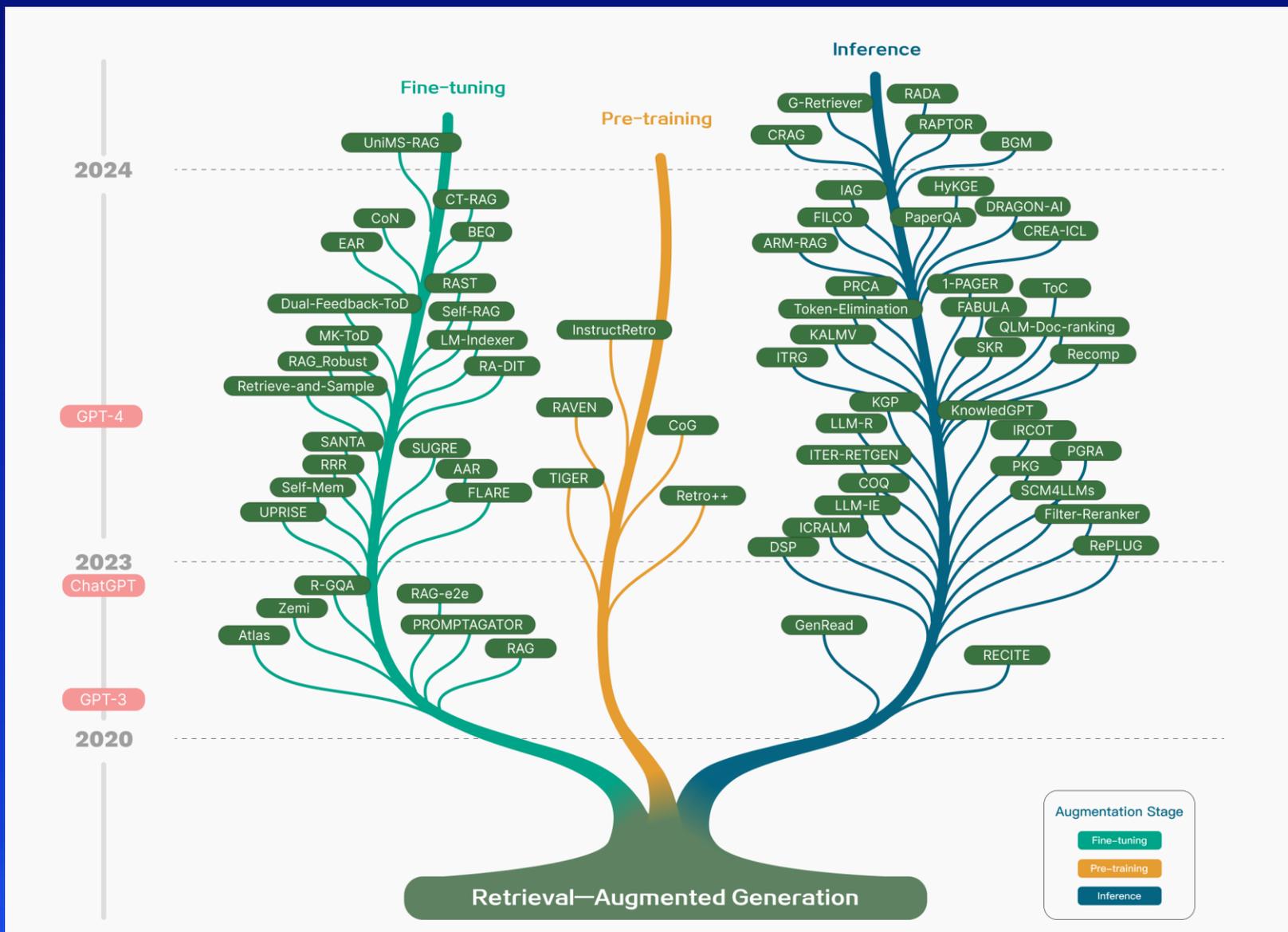
► 推荐的优化路径

Prompt engineering -> RAG -> fine-tuning



OpenAI: A Survey of Techniques for Maximizing LLM Performance

▶ 两条路径: Finetune vs. RAG



▶ LLM推理成本和性能优化：微调

- **模型的参数量决定推理使用的资源和成本**

- $GPU\ Memory = \frac{(Parameter * 4bytes)}{(32/QuantizationBit)} * 1.2$

- **模型的输入和输出大小影响推理的成本和性能**

- Prompt的复杂度影响首token返回时延 (Time To First Token, TTFT)

- **微调的作用：**

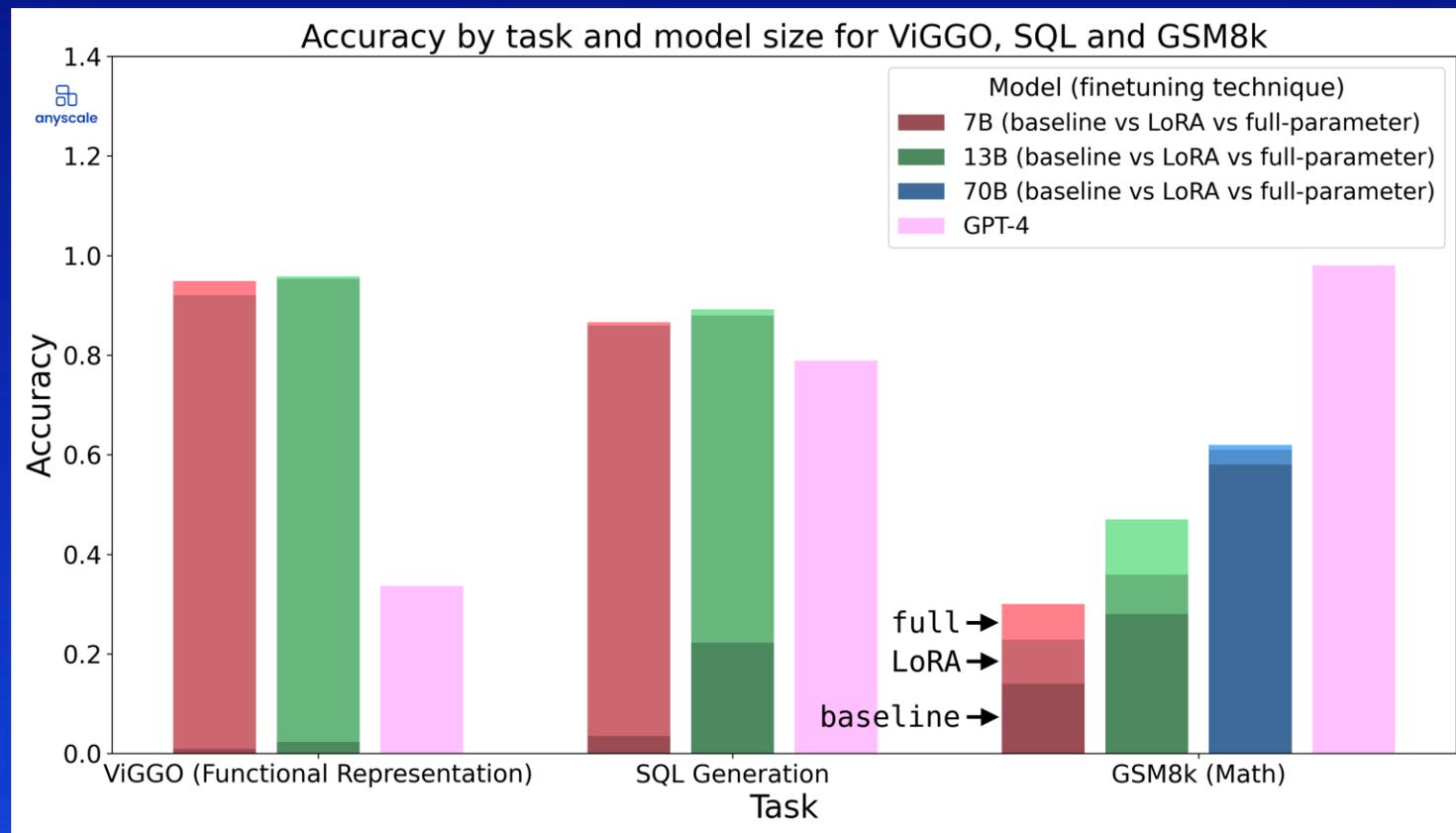
- 优化Prompt的输入、降低延迟、使用更小的模型完成专门的任务

OpenAI API Pricing

Model	Input	Output
gpt-4-0125-preview	\$10.00 / 1M tokens	\$30.00 / 1M tokens
gpt-4-1106-preview	\$10.00 / 1M tokens	\$30.00 / 1M tokens
gpt-3.5-turbo-0125	\$0.50 / 1M tokens	\$1.50 / 1M tokens
gpt-3.5-turbo-instruct	\$1.50 / 1M tokens	\$2.00 / 1M tokens

▶ 模型微调的效果

- 在“简单”的数据抽取和格式对齐任务中，小模型微调后能够达到大模型的性能（效果）。
- 在“复杂”的任务中，大模型的参数量是模型性能（效果）的基础。



▶▶ LLM应用效果优化：微调 vs. RAG

- **微调能做的：**

- 添加静态、私有的数据，优化模型在领域场景中的性能，减少模型幻觉
- 使用小模型，优化推理成本和效率
- 优化模型的输入输出，减少延迟，降低成本

- **RAG能做的：**

- 引入新的信息，模型训练没有见过的信息。
- 要求模型根据检索获得的数据回答，减少幻觉。

- **RAG不能做的：**

- 教模型学会一个广泛领域的知识，例如医学，法律
- 教会模型学会一门新的语言、格式、或是风格

二者的对比类似“开卷考试” vs “考前刷题复习”
并非是非此即彼，可以是相互配合的关系。

PART 03

效果优化的工具

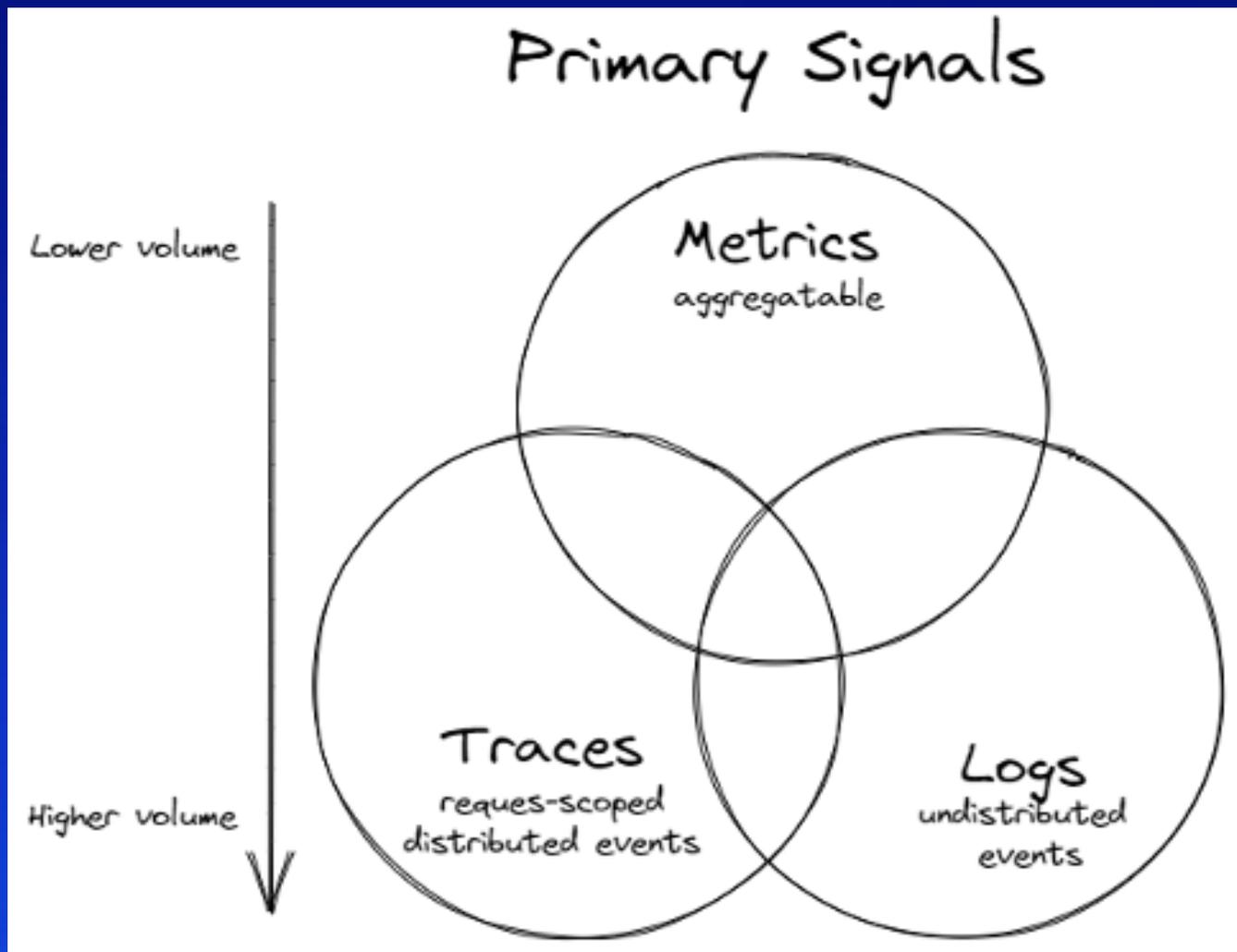
▶▶ 选择合适的模型：评测

不同基础模型的对比

同一模型不同微调版本的对比

同一模型不同量化版本的对比

► 提高问题排查效率：可观测性



▶▶ RAG应用的开发流程 (部分)



▶▶ 如何debug一个复杂链路：Tracing

每个步骤的输入输出是什么？

每个步骤的耗时是多少？

每次LLM调用消耗的token是多少？

▶ Tracing case study: Arize AI Phoenix

Trace Details

Trace Status: OK Latency: 3.84s Evaluations: Hallucination factual QA Correctness correct

```
graph TD; query[query chain 3.84s] --> retrieve[retrieve retriever 0.44s]; query --> embedding[embedding embedding 0.27s]; query --> synthesize[synthesize chain 3.40s]; synthesize --> llm[llm llm 653 3.39s];
```

retrieve

retriever retrieve 0.44s

Info Evaluations 0 Attributes Events 0

Input

How do I log a prediction using the python SDK?

Documents Relevance ndcg 1.00 Relevance precision 1.00 Relevance hit true

document 068d4de8-7338-4a57-9084-d371f4482806 score 0.82

```
schema = Schema( prediction_id_column_name="prediction_id", #needs to be the same as above actual_label_column_name="actual_label", ... ) `` `{% endtab %} {% tab title="Python Single Record" %} To log delayed actuals using the Python Single Record SDK, simply match the actual `prediction_id` with its corresponding prediction. From there, Arize will automatically identify the join and match the data together.&#x20; ``python #log the features & prediction response = arize.log( prediction_id='pLED4eERDCasd9797ca34', model_id='sample-model-1', model_type=ModelTypes.SCORE_CATEGORICAL, environment=Environments.PRODUCTION, model_version='v1', prediction_timestamp=1618590882, features=features, prediction_label=('Fraud',.4), tags=tags ) #log the actual actual_response = arize.log( prediction_id='pLED4eERDCasd9797ca34', model_id='sample-model-1', model_type=ModelTypes.SCORE_CATEGORICAL, environment=Environments.PRODUCTION, actual_label=('Fraud',1), tags=tags) `` `{% endtab %} {% endtabs %}
```

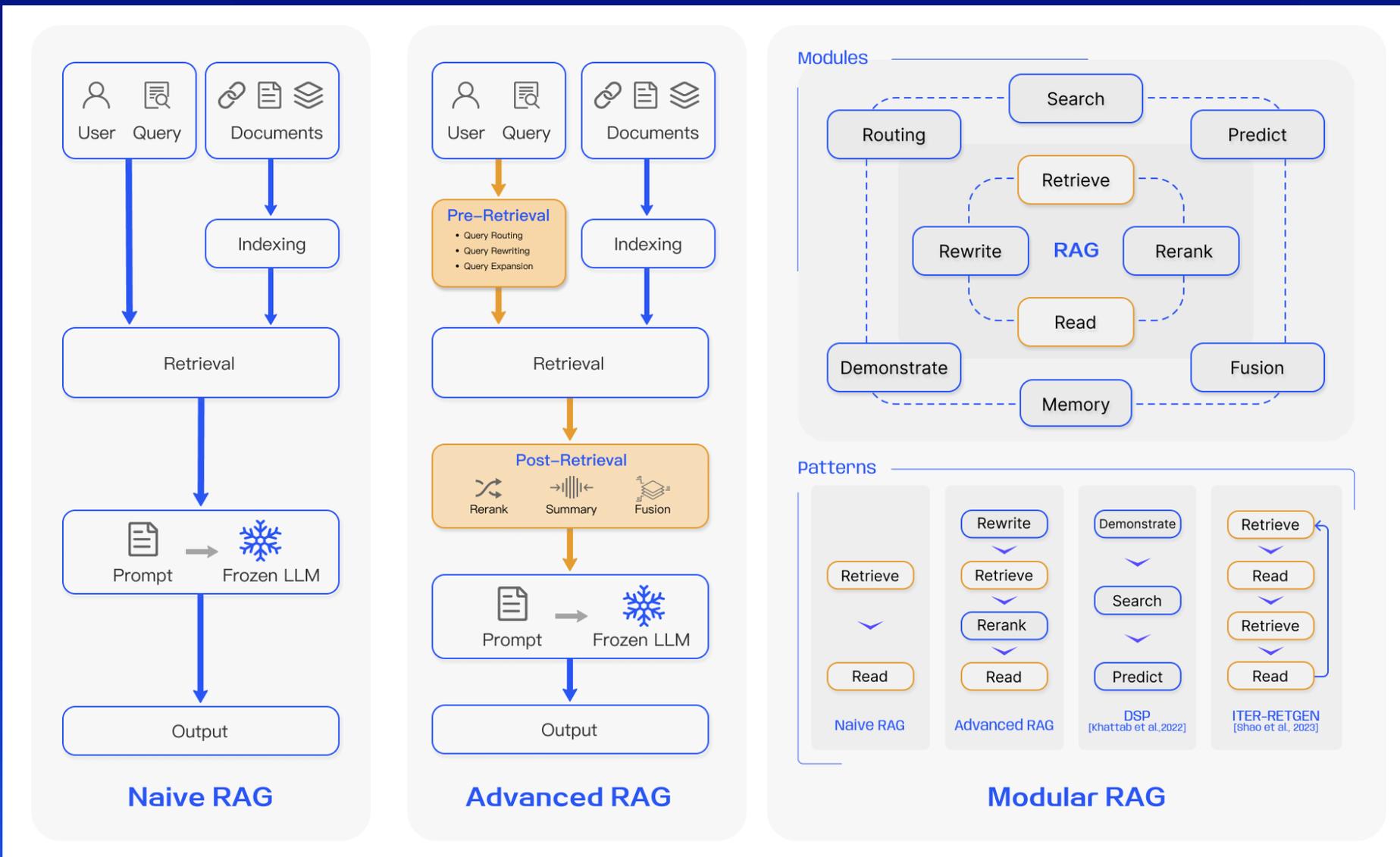
Evaluations

Relevance relevant score 1.00

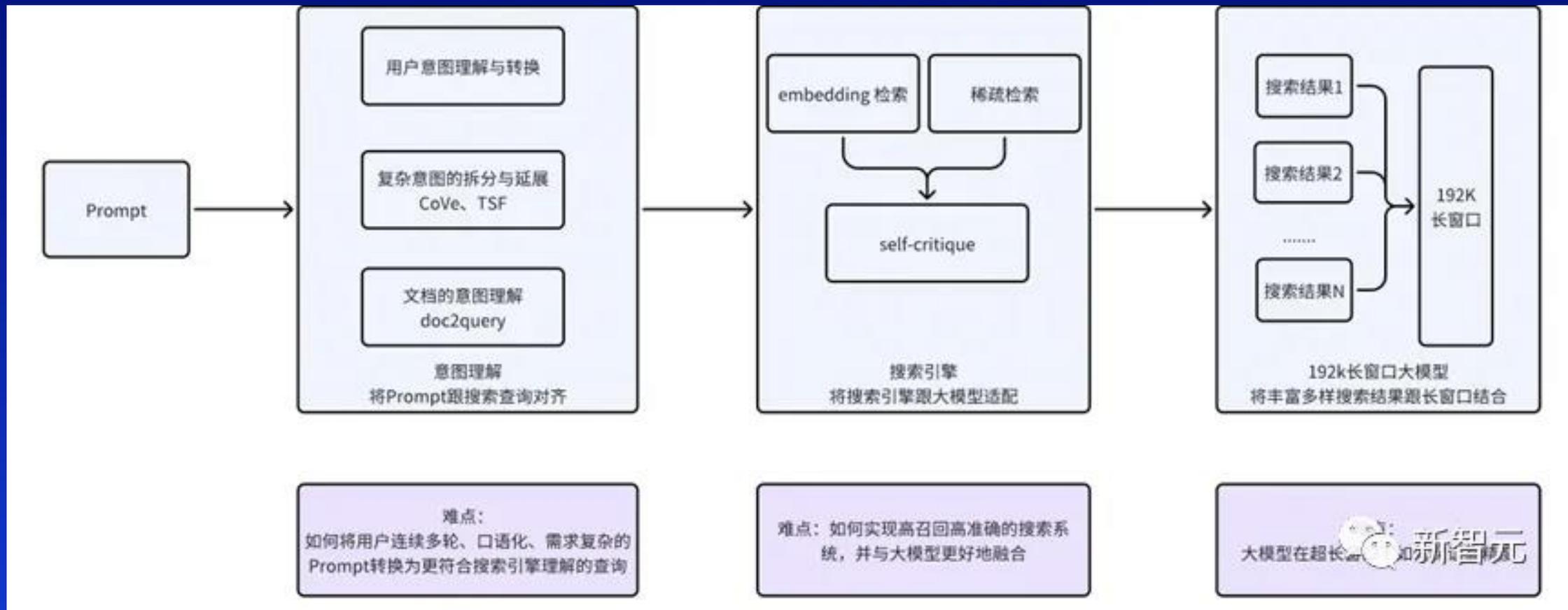
▶▶ 如何系统的“炼丹”：实验管理



▶ RAG效果调优路径



百川RAG: 长窗口模型 + 搜索增强



难点1: 如何让搜索引擎理解用户复杂的提示

难点2: 如何实现适配大语言模型的搜索系统

PART 04

阿里云AI原生应用开发实践

▶ PAI产品架构

场景化
解决方案

AI应用：自动驾驶 / 科研智算 / 金融风控 / 智能推荐 / 智能设计 / 智慧城市 / 智能制造 / 智慧医疗 / 智慧法务 / ...

模型服务
(MaaS)

ModelScope 魔搭社区

PAI-DashScope 模型服务灵积

第三方MaaS平台

灵骏智算服务
&
机器学习框架
(PaaS)

快速开始：PAI-QuickStart / PAI-智码实验室 / PAI-DSW Gallery

工作空间

标注服务
PAI-iTAG

特征平台
PAI-Feature
Store

可视化建模
PAI-Designer

交互式建模
PAI-DSW

分布式训练
PAI-DLC

模型在线服务
PAI-EAS

开发者工具
CLI / PaiFlow /
OpenAPI

权限管理

MLOps

AI资产管理（数据集 / 模型 / 镜像 / 代码 / 自定义组件 / ...）

云产品依赖

优化与加速（DatasetAcc 数据集加速 / TorchAcc 训练加速 / EPL 并行训练框架 / Blade推理加速 / AI Master 自动容错训练 / EasyCkpt 秒级异步训练快照）

机器学习框架（PAI-TensorFlow / PAI-PyTorch / Alink / Spark, EasyRec / EasyPhoto / EasyTransfer / Megatron / DeepSpeed / RLHF）

计算资源
&
基础设施
(IaaS)

PAI-灵骏计算资源

云原生通用计算资源

大数据计算资源（MaxCompute / EMR / Flink）

异构计算磐久服务器（CPU、GPU）

高速 RDMA 网络（RoCE）

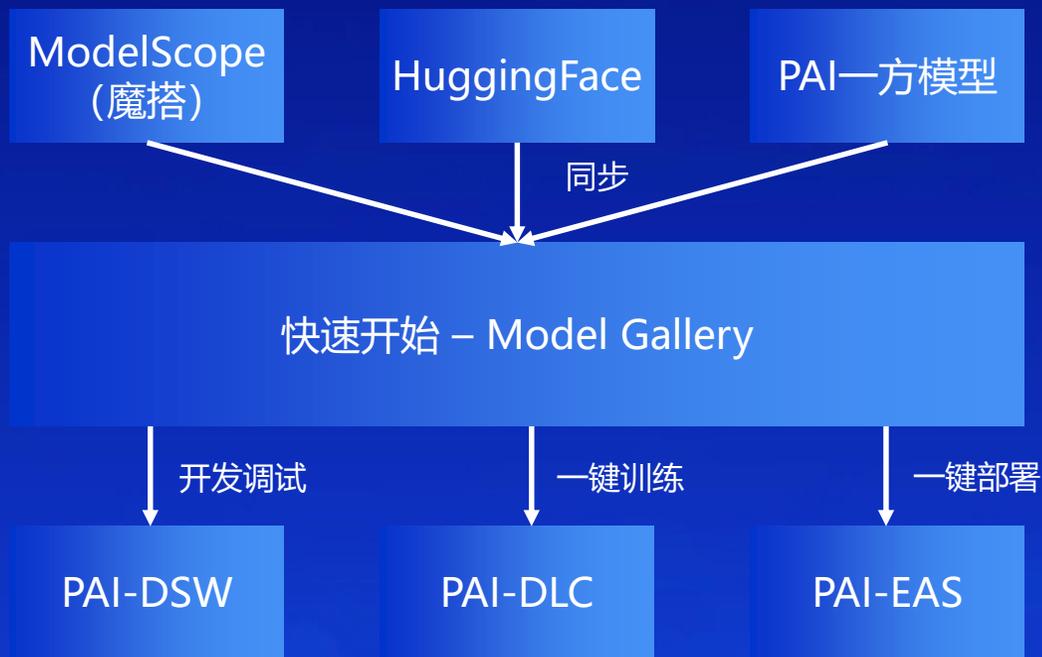
分布式存储 CPFS/NAS/OSS

容器服务 ACK

弹性计算 ECS

低PUE液冷/风冷，模块化IDC设施

▶▶ 以模型为中心的MLOps/LLMOps平台：PAI-QuickStart



- 一站式模型训练->部署的全链路体验
- 内置算法工程优化, 提升迭代效率
- 结合平台优化能力, 提供极致的性能和性价比

▶ PAI-QuickStart模型微调工具链

人工智能平台PAI / 快速开始 / 模型列表 / qwen1.5-7b-chat

← qwen1.5-7b-chat

生成式AI/大语言模型 Chinese PyTorch ModelScope

参数量 7B 04/16 2024

Qwen1.5-7B-Chat大语言模型

模型简介

通义千问1.5-7B-Chat (Qwen1.5-7B-Chat) 是阿里云研发的通义千问大模型系列的70亿参数规模的模型。Qwen1.5-7B是基于Transformer的大语言模型，在超大规模的预训练数据上进行训练得到。预训练数据类型多样，覆盖广泛，包括大量网络文本、专业书籍、代码等。同时，在Qwen1.5-7B的基础上，我们使用对齐机制打造了基于大语言模型的AI助手Qwen1.5-7B-Chat。

本模型可以直接部署，直接部署的模型采用Qwen1.5-7B-Chat作为预训练模型，可以根据用户提供的任意文本进行续写。

训练数据格式

训练数据接受Json格式输入，每条数据由问题、答案组成，分别用"instruction"和"output"字段表示，例如：

```
[
  {
    "instruction": "你是一个心血管科医生，请根据患者的问题给出建议：我患高血压五六年啦，天天吃药吃烦啦，哪种东西能根治高血压，高血压克星是什么？",
    "output": "高血压的患者可以吃许多新鲜的水果蔬菜或者是芹菜山药之类的食物，可以起些降血压的作用，另外高血压的患者平时也应当注意低盐，低脂，低胆固醇饮食，适当的实施体育运动和锻炼高血压的患者"
```

帮助文档

部署 微调训练 评测 PAI SDK

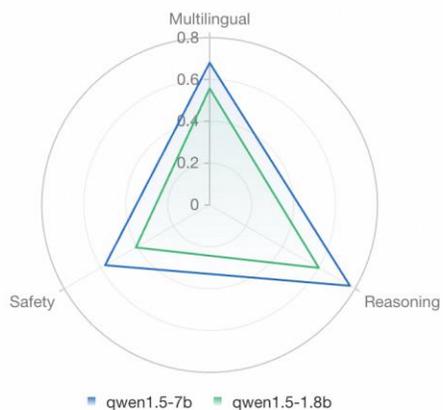
QuickStart – 模型评测

人工智能平台PAI / 模型评测 / 评测结果对比

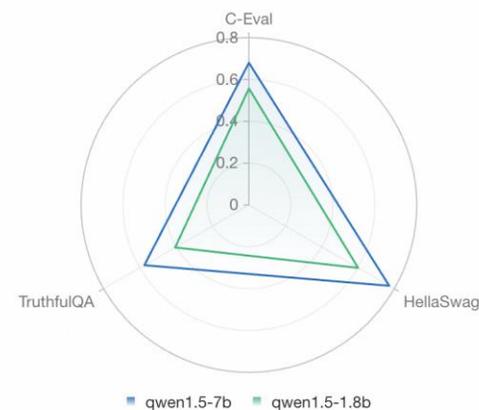
← 评测结果对比

自定义数据集评测结果 公开数据集评测结果

各领域得分情况



各数据集得分情况



导出



任务名称	模型名称	C-Eval	TruthfulQA	HellaSwag
qwen1.5-7b	qwen1.5-7b-chat	0.680	0.576	0.772
qwen1.5-1.8b	qwen1.5-1.8b-chat	0.557	0.406	0.601

阿里云 工作台 华东2 (上海)

人工智能平台PAI / 快速开始 / 模型列表 / qwen1.5-7b-chat

← qwen1.5-7b-chat

生成式AI/大语言模型 Chinese PyTorch ModelScope

参数量 7B 04/16 2024

Qwen1.5-7B-Chat大语言模型

模型简介

通义千问1.5-7B-Chat (Qwen1.5-7B-Chat) 是阿里云研发的通义千问大模型预训练数据类型多样, 覆盖广泛, 包括大量网络文本、专业书籍、代码等。本模型可以直接部署, 直接部署的模型采用Qwen1.5-7B-Chat作为预训练数据。

训练数据格式

训练数据接受Json格式输入, 每条数据由问题、答案组成, 分别用"instruction"和"output"表示。

```
[
  {
    "instruction": "你是一个心血管科医生, 请根据患者的问题给出建议。",
    "output": "高血压的患者可以吃许多新鲜的水果蔬菜或者是芹菜山药"
  }
]
```

PAI-快速开始中展示模型/数据集/文件均来自于第三方, 阿里云不保证合规性, 请您在使用前慎重阅读《阿里云服务专用条款协议》及页面展示信息等。如您发现任何模型/数据集/文件有问题, 请及时联系我。

微调训练

使用默认数据或自己的业务数据进行模型训练, 从而得到您的专属场景模型。

训练设置

* 任务名称

最大运行时长

 分钟

为0则不会因超时中断任务, 停止后不可恢复

数据集配置

* 训练数据集

默认路径

+ 添加验证数据集

实验配置

训练输出配置

计算资源配置

* 资源组类型

训练

QuickStart – 实验管理

Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing: 0.6

Horizontal Axis: STEP (1)

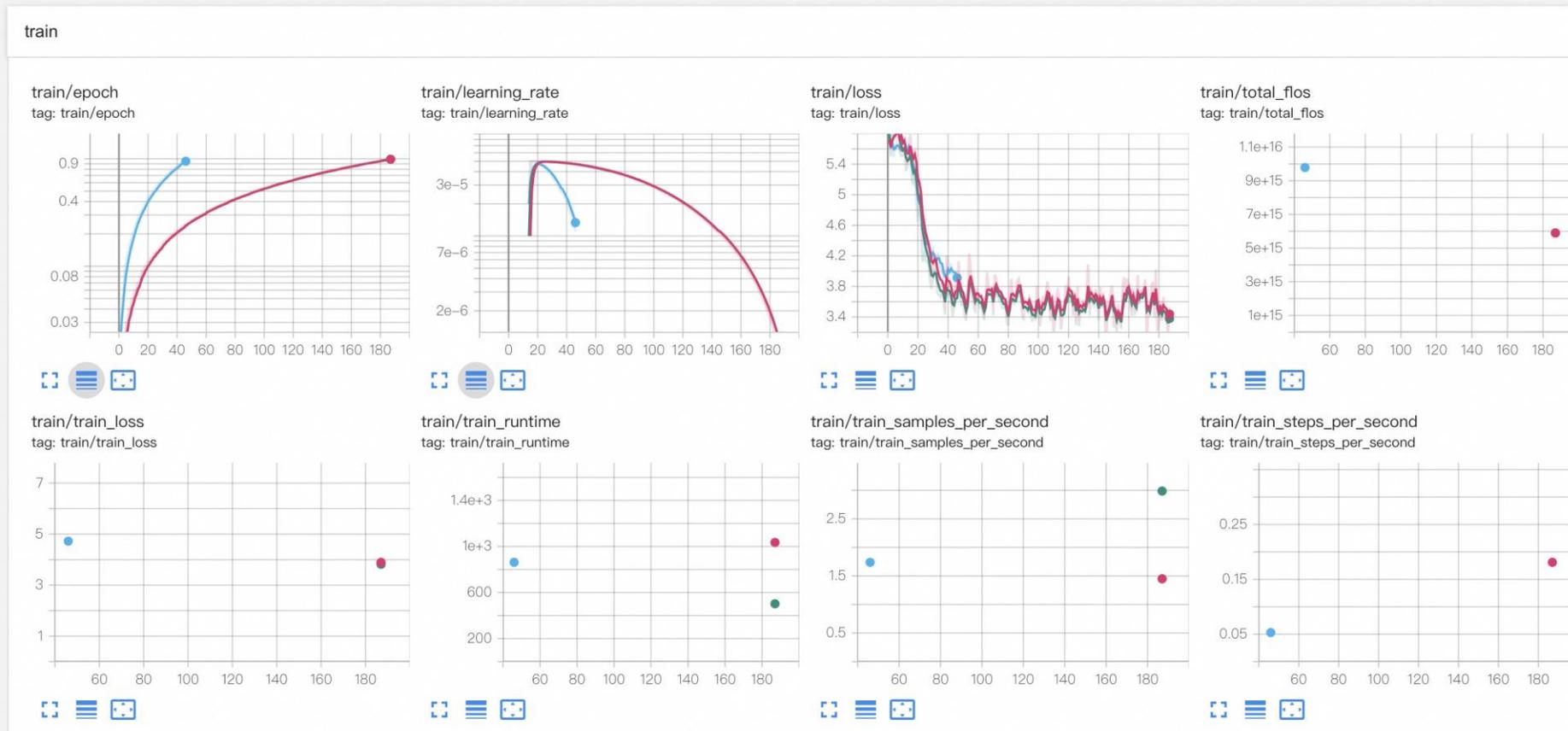
Runs

Write a regex to filter runs

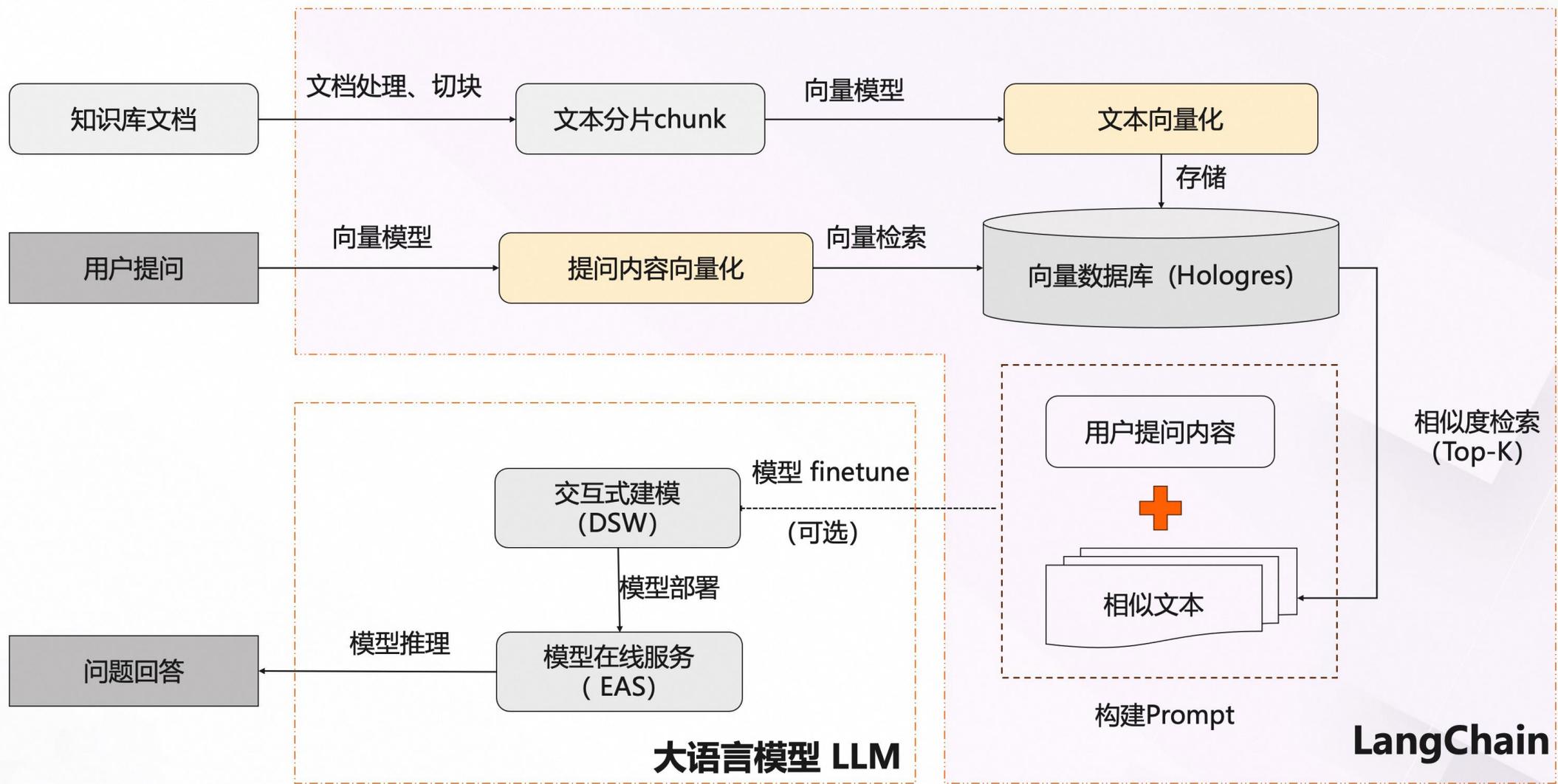
- /train1knnk6potc
- /train1qwgw6rdje
- /train74lhja55f95

multiFS:///tmp/tb_logdir_1

Filter tags (regular expressions supported)



▶ PAI RAG最佳实践



▶ PAI RAG最佳实践

Settings Upload Chat

Which query do you want to use?

Vector Store LLM

Langchain(Vector Store + LLM)

Parameters of Vector Retrieval

Top K (choose between 0 and 100)

Similarity Distance Threshold (The more similar the vectors, the smaller the value.)

Re-Rank Model

No Re-Rank

BGE-Reranker-Base

BGE-Reranker-Large

Keyword Retrieval

Embedding Only

Keyword Ensembled

Chatbot

Enter your question.

Submit Summary Clear History Clear

▶ PAI LLM推理性能优化：BladeLLM

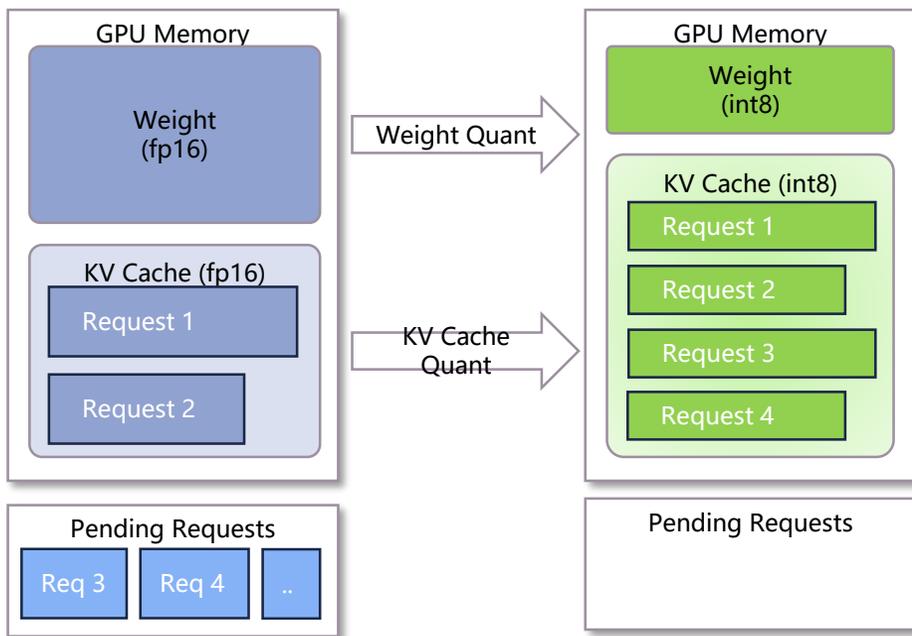
平台和基础引擎视角，部署优化需要关注：

- 基础性：核心优化技术的高效实现作为性能基础
- 综合性：完善的应用接口和多样的优化功能支持
- 规模化可扩展性：突破常规限制，面向更大规模性能需求



▶ 基础性能优化：量化压缩

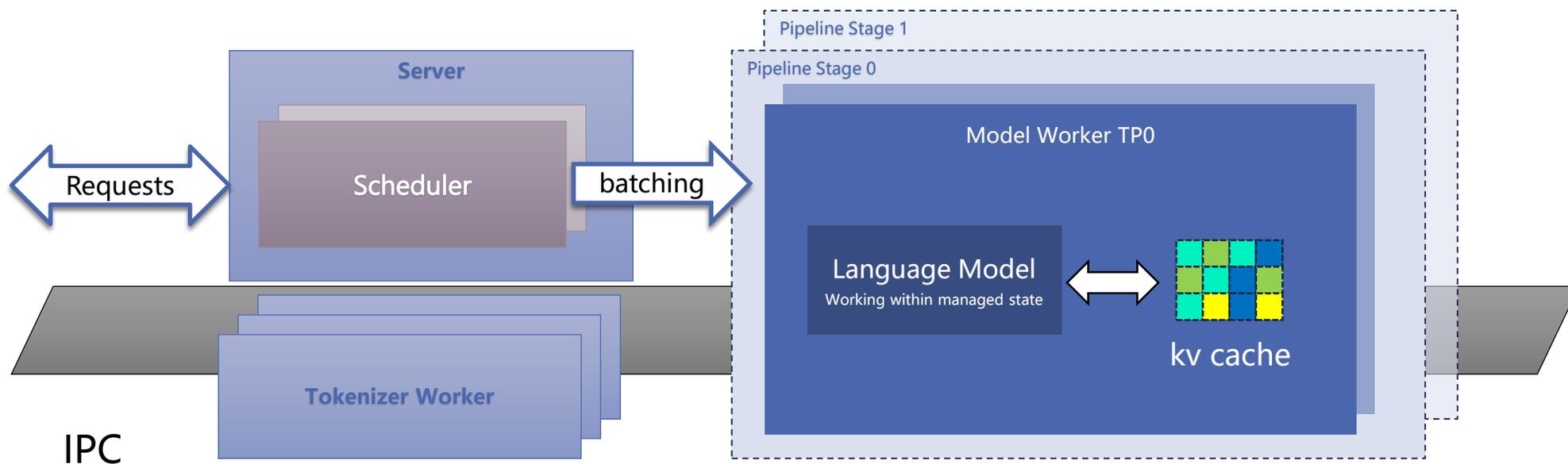
- 多样的量化算法策略和高效的量化算子，以适应调优需求提升精度和性能。



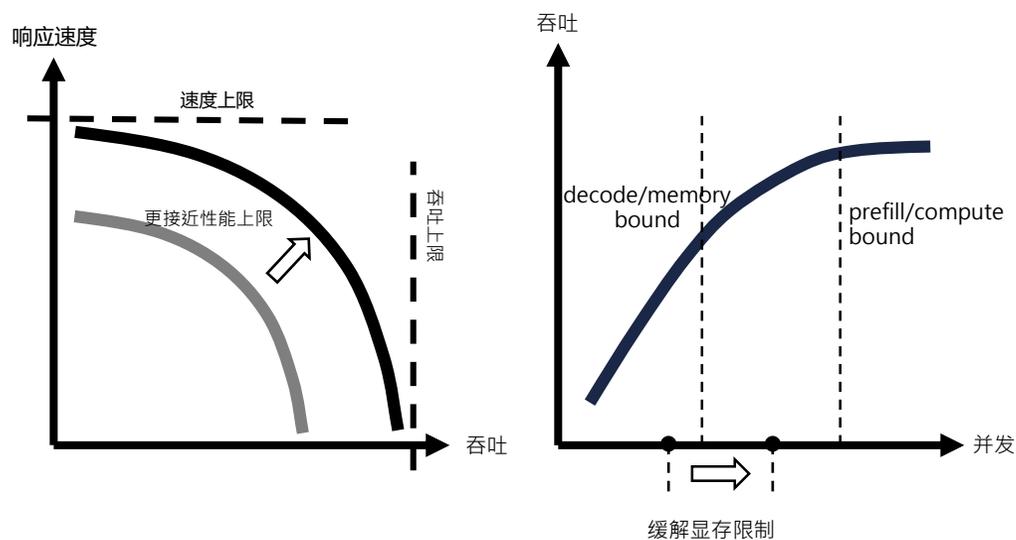
Quantization Mode	DType	Granularity	Mixed Precision
Weight-Only	int8/int4	per-channel/sub-channel	y
Activation-Weight	fp8/int8	per-channel/sub-channel	y
KV Cache	fp8/int8/int4	per-head/sub-head	y

▶ 基础性能优化：高效进行时

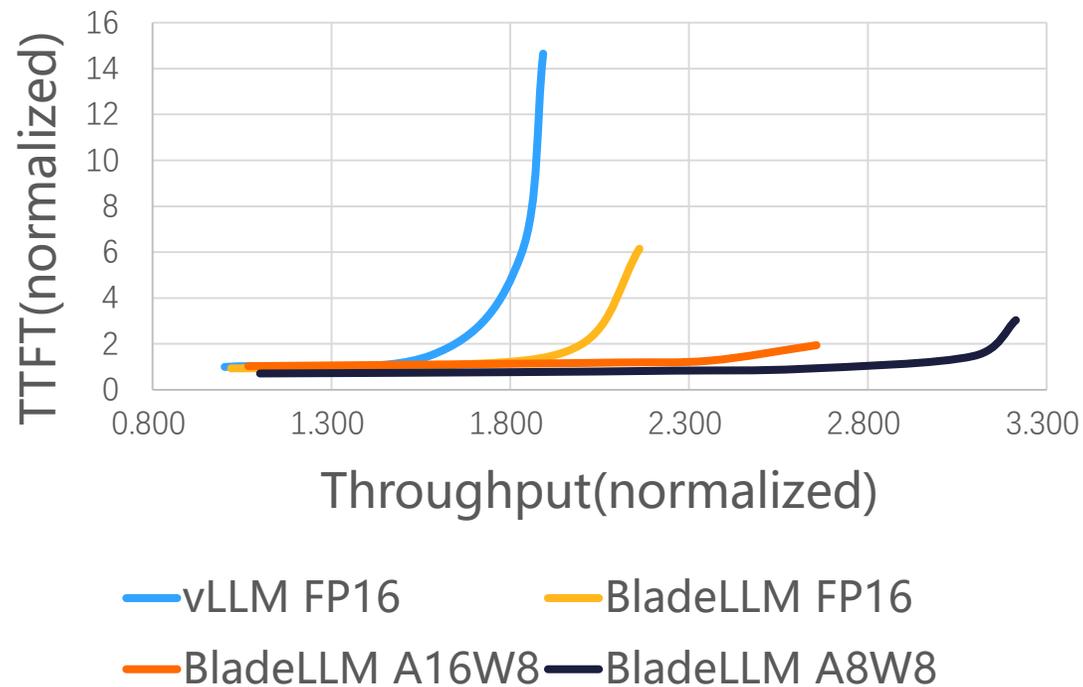
- 服务架构：server/model/tokenizer异步执行；TP/PP并行策略；
- 请求调度：支持多种batching和优先级调度策略；
- 运行时优化：Native Runtime, CUDA Graph等多种执行模式。



基础性能优化收益分析



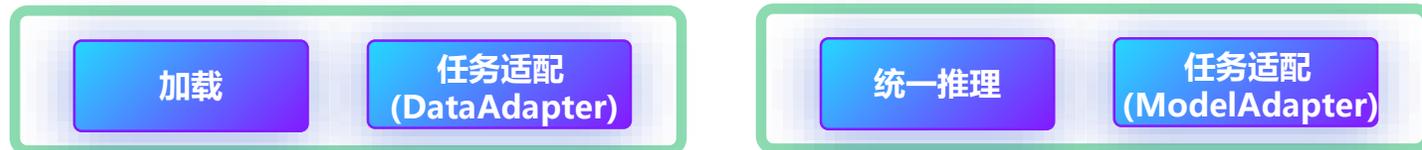
- 通过基础优化接近性能上限和缓解显存瓶颈



- 实际场景的性能提升示例

LLM评测框架：Eval-Scope

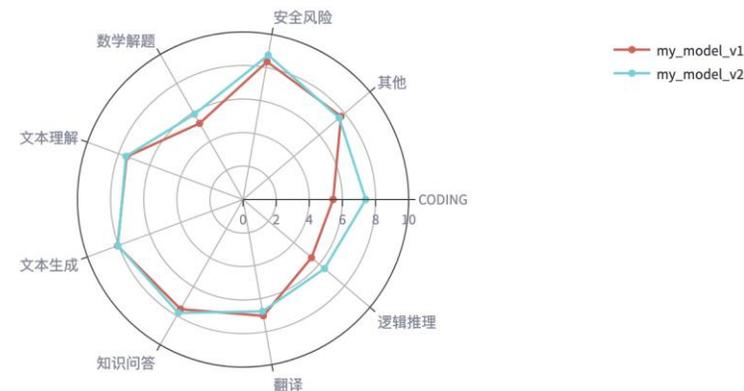
Eval-Scope是PAI与ModelScope共建的一款开源的轻量化LLM评测工具



模型分类别得分

	类别	Case数量	my_model_v1	my_model_v2	分差百分比
0	安全风险	10	8.35	8.75	4.79%
1	文本生成	94	8.06	8.04	-0.33%
2	其他	19	7.72	7.58	-1.87%
3	知识问答	66	7.55	7.81	3.46%
4	文本理解	57	7.48	7.53	0.70%
5	翻译	11	7.05	6.77	-3.87%
6	CODING	16	5.44	7.41	36.21%
7	逻辑推理	24	5.40	6.42	18.92%
8	数学解题	20	5.25	5.88	11.90%
9	总计	317	7.29	7.56	3.66%

模型得分雷达图



科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **上海站**

K+ 全球软件研发行业创新峰会

时间: 2024.06.21-22

 **K+峰会**  **敦煌站**

K+ 思考周®研习社

时间: 2024.10.17-19

 **K+峰会**  **香港站**

K+ 思考周®研习社

时间: 2024.11.10-12



K+峰会详情



 **AiDD峰会**  **上海站**

AI+研发数字峰会

时间: 2024.05.17-18

 **AiDD峰会**  **北京站**

AI+研发数字峰会

时间: 2024.08.16-17

 **AiDD峰会**  **深圳站**

AI+研发数字峰会

时间: 2024.11.08-09



AiDD峰会详情



THANKS

