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 + 技术团队的共同选择

KEYLINK ing

K+峰会

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

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

演讲嘉宾

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 ）

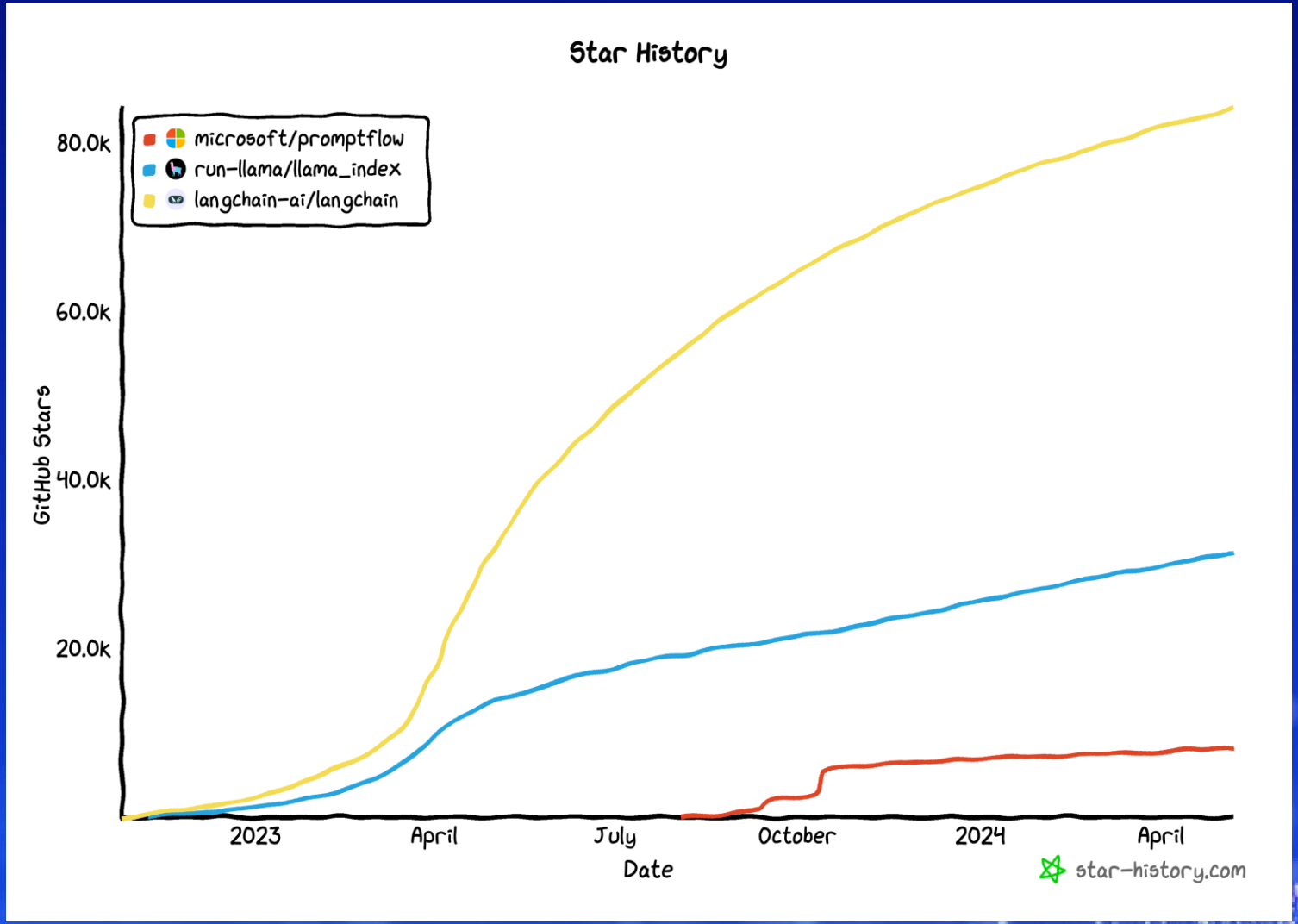
55.1%

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


Star History
microsoft/promptflow
run-llama/llama_index
langchain-ai/langchain
80.0k
60.0k
40.0k
20.0k
GitHub Stars
2023
April
July
October
2024
April
Date
star-history.com

典型的AI应用开发全流程

AI原生应用开发的挑战

成本	效率	效果
• 训练成本 • 推理成本	• 迭代次数多 • Debug困难	• 幻觉 • 内容合规

PART 02

成本、性能和效果的综合考量：微调和RAG

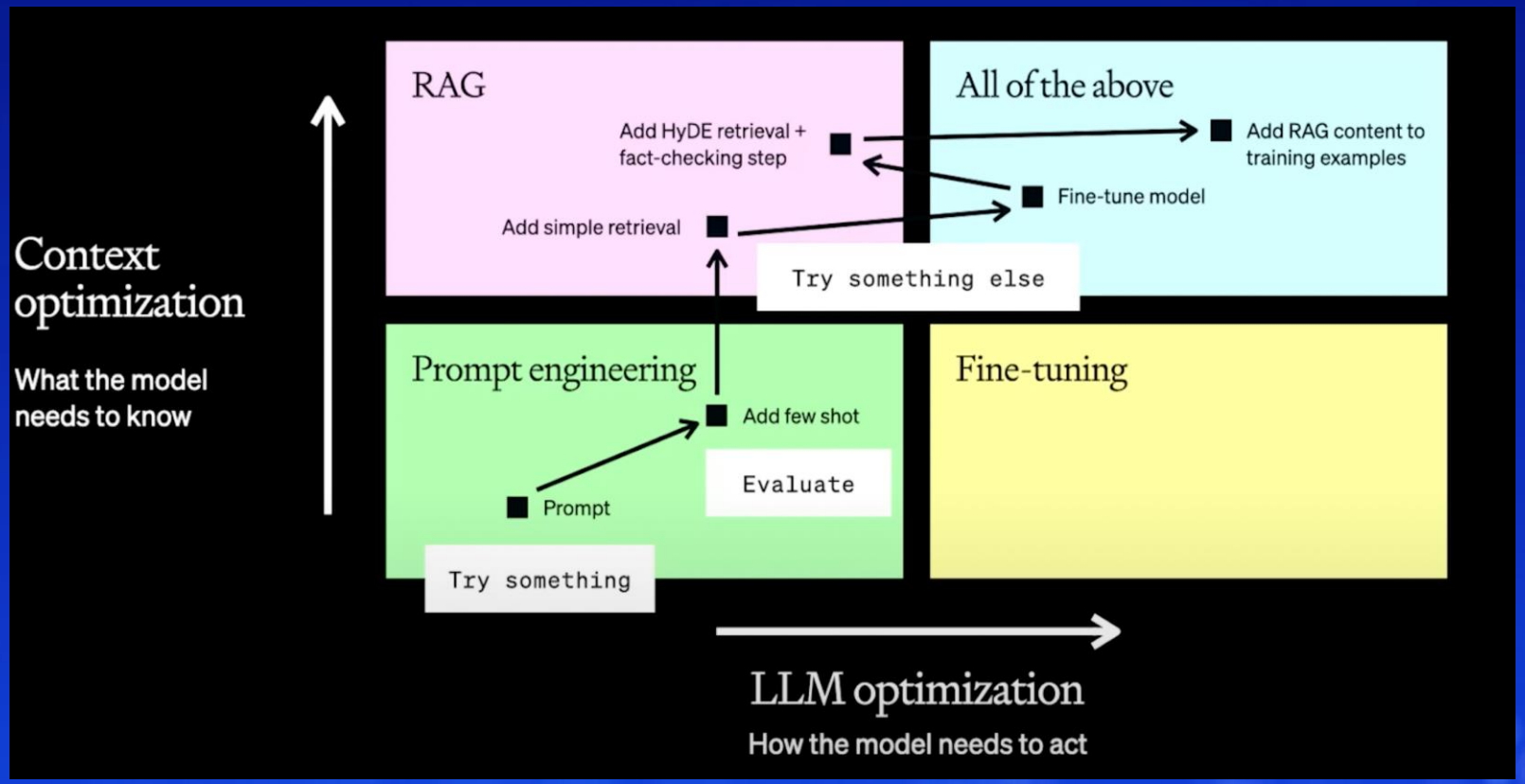
LLM的应用面临什么挑战

- **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

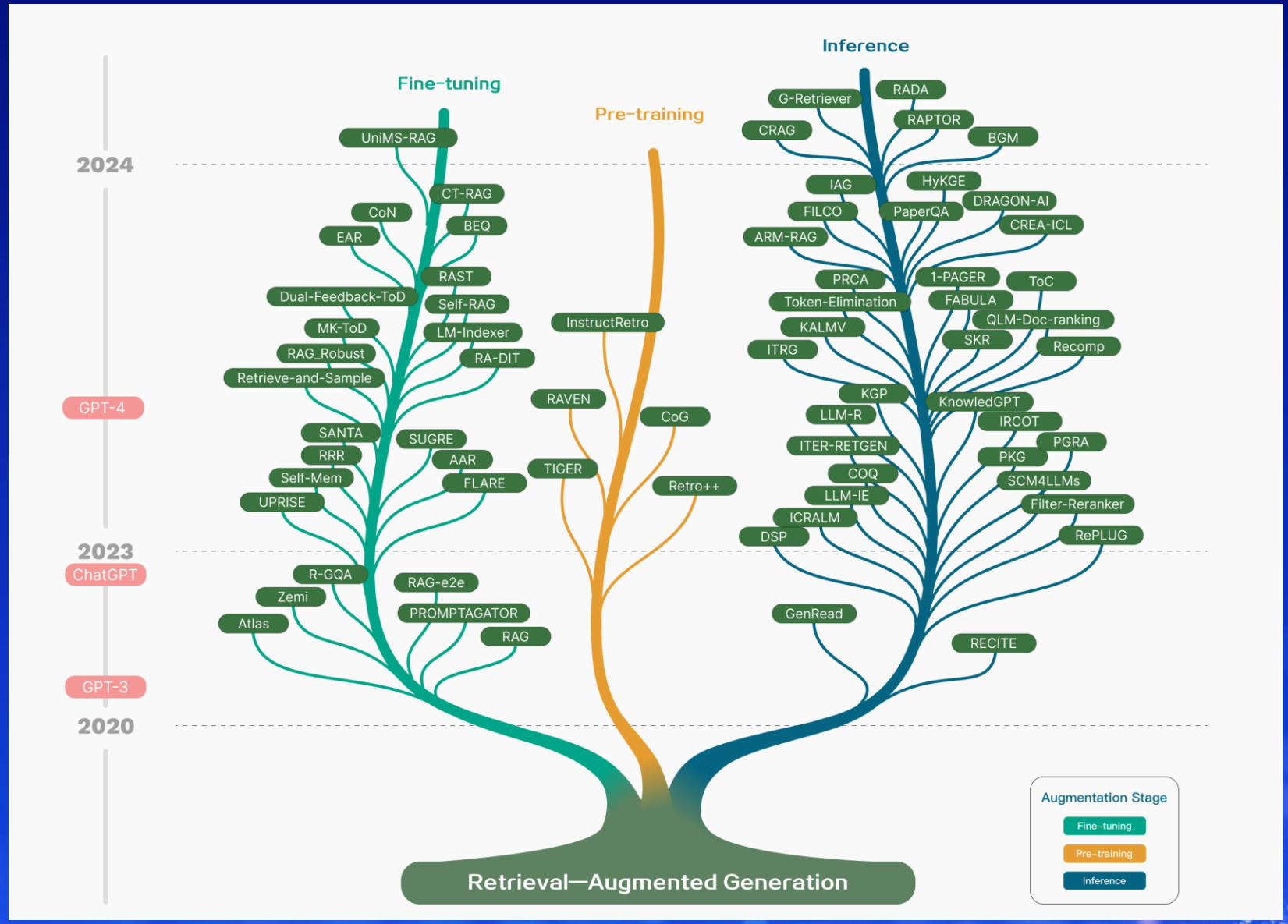
推荐的优化路径

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

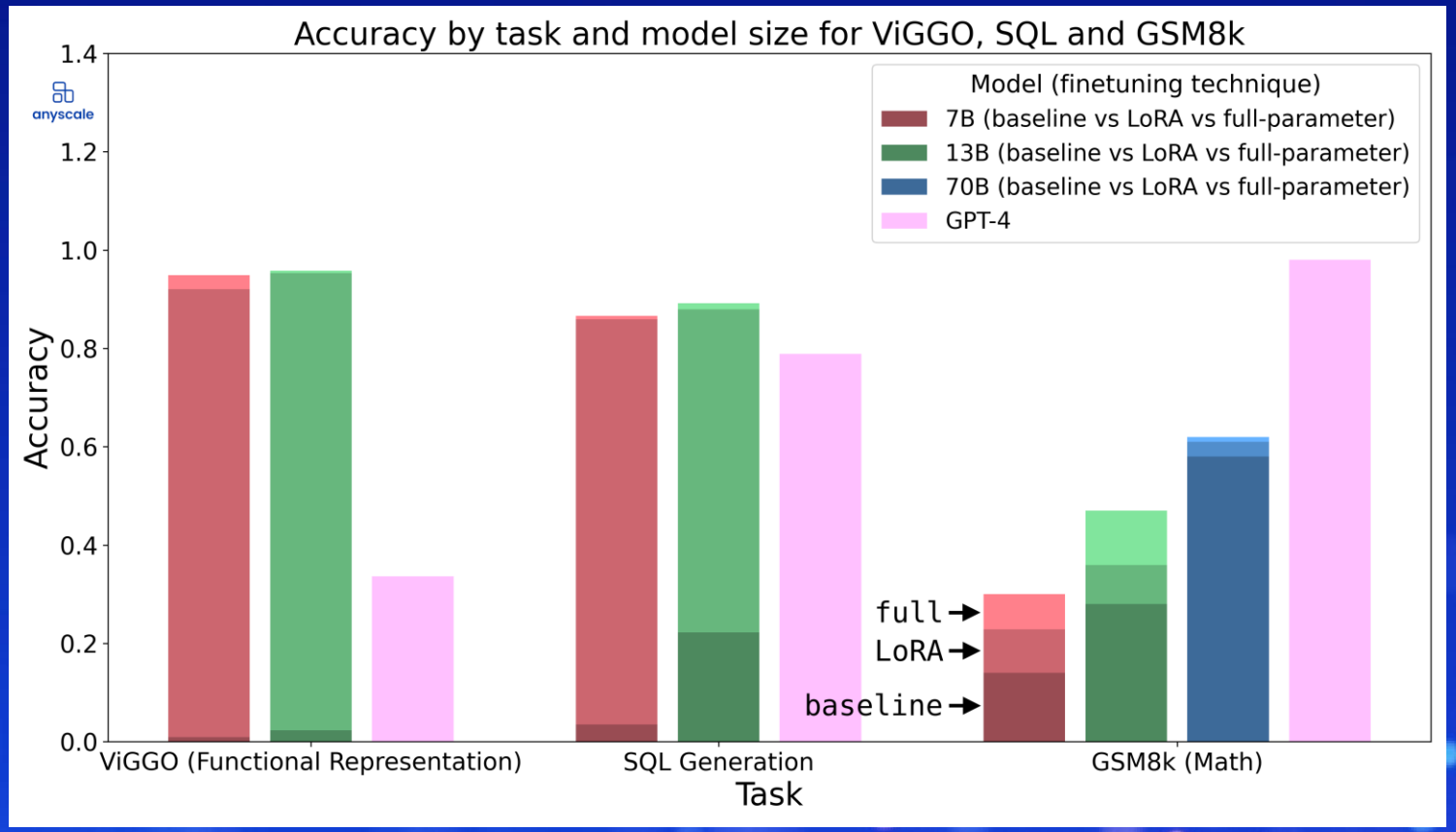
- **模型的参数量决定推理使用的资源和成本**
  - $GPU\ Memory = \frac{(Parameter * 4bytes)}{(32/QuantizationBit)} * 1.2$
- **模型的输入和输出大小影响推理的成本和性能**
  - Prompt的复杂度影响首token返回时延（Time To First Token，TTFT）
- **微调的作用：**
  - 优化Prompt的输入、降低延迟、使用更小的模型完成专门的任务

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。



https://www.anyscale.com/blog/fine-tuning-llama-2-a-comprehensive-case-study-for-tailoring-models-to-unique-applications

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”

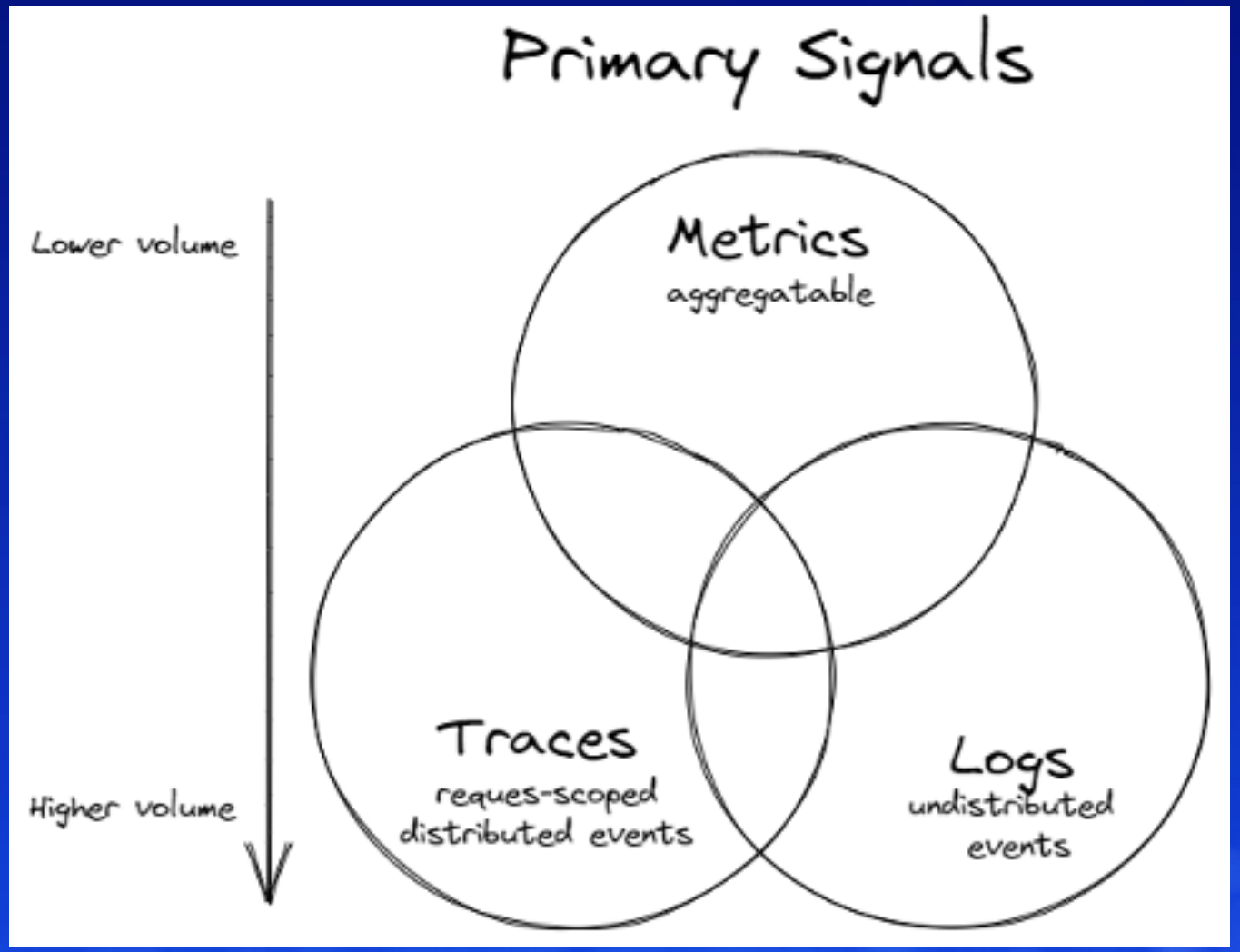
并非是非此即彼，可以是相互配合的关系。

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

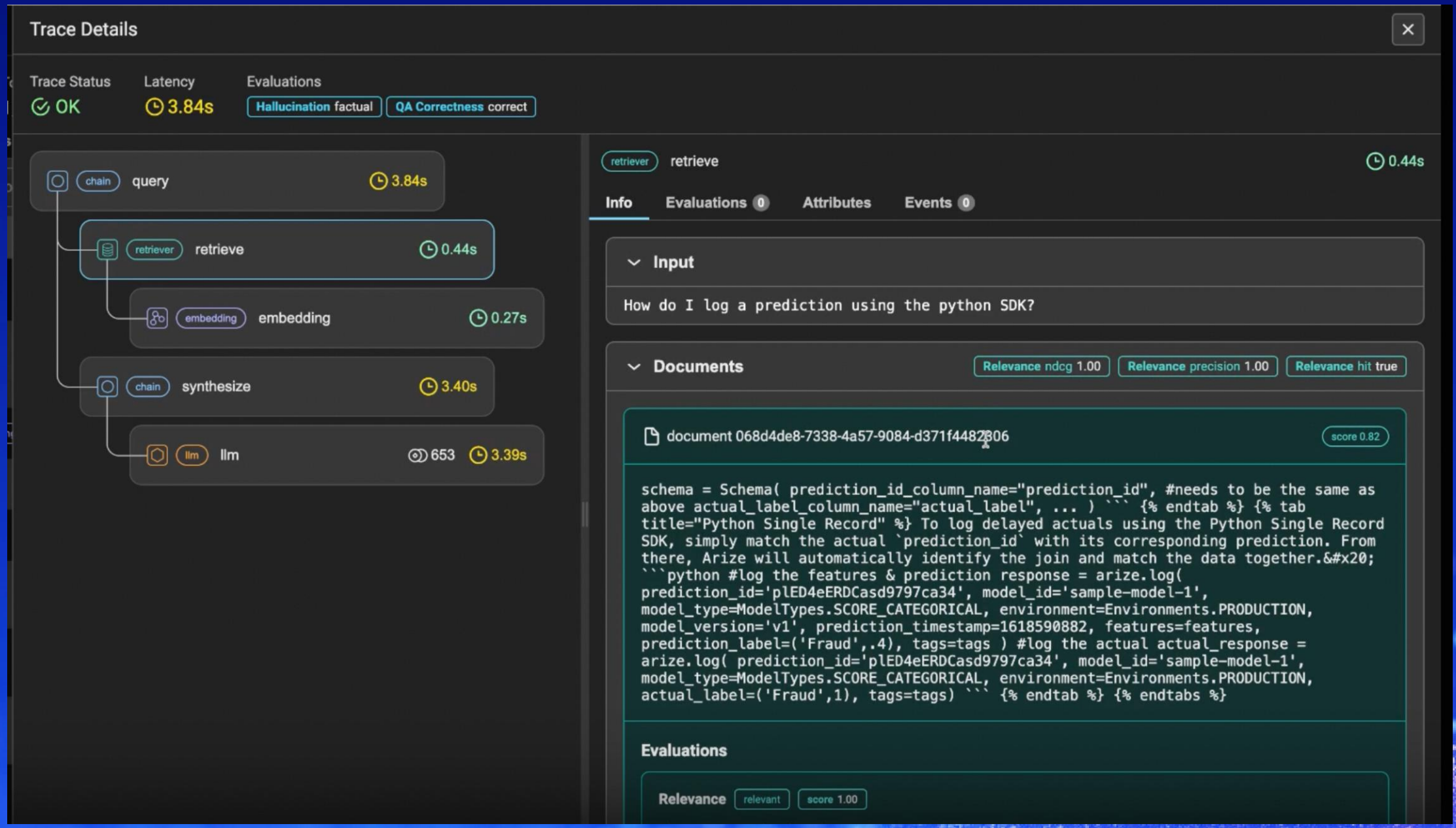


RAG应用的开发流程（部分）

如何debug一个复杂链路：Tracing

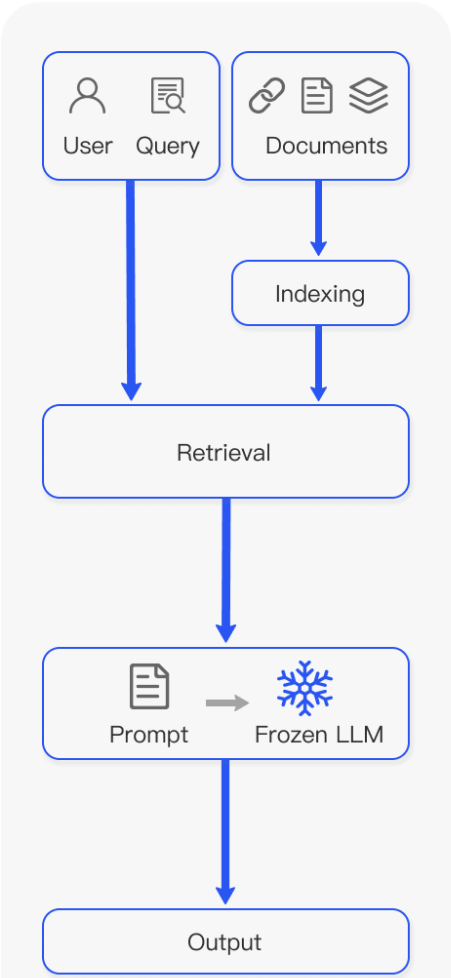
- 每个步骤的输入输出是什么?
- 每个步骤的耗时是多少?
- 每次LLM调用消耗的token是多少?

Tracing case study: Arize AI Phoenix

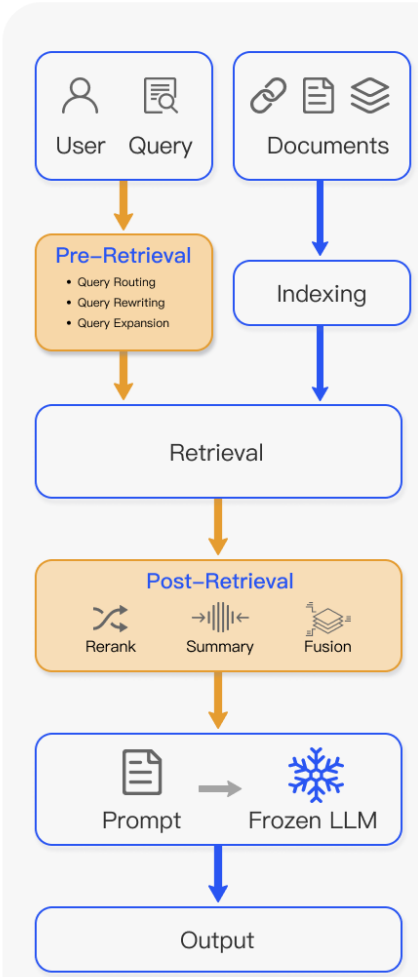


如何系统的“炼丹”：实验管理

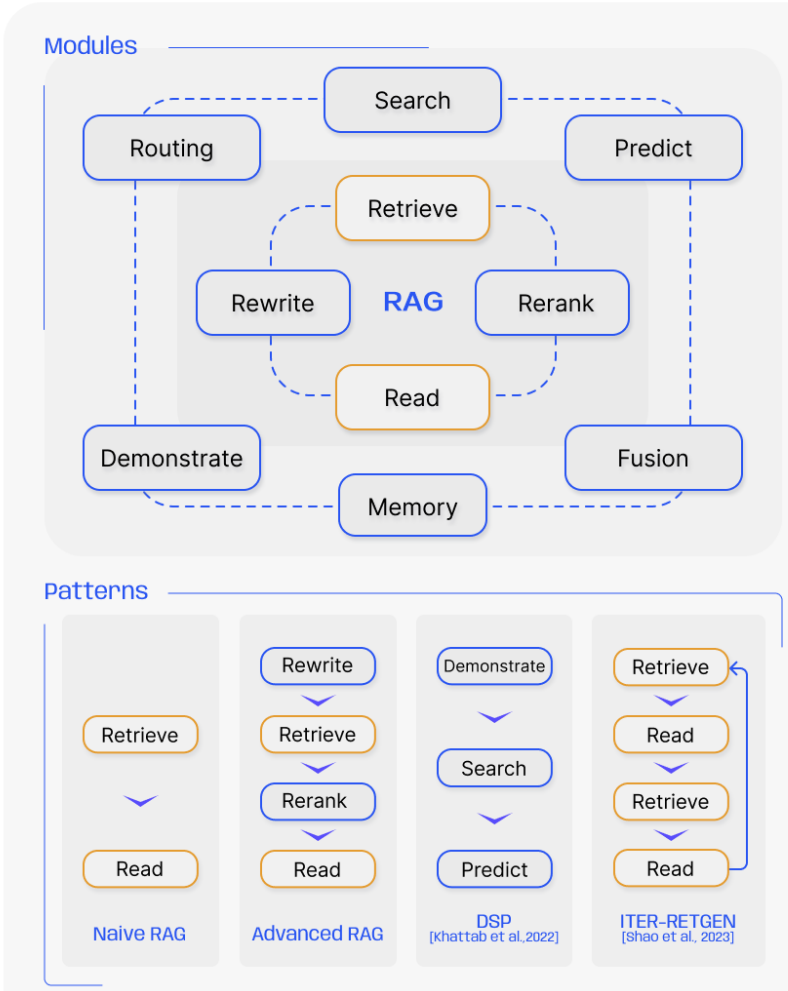
RAG效果调优路径



Naive RAG

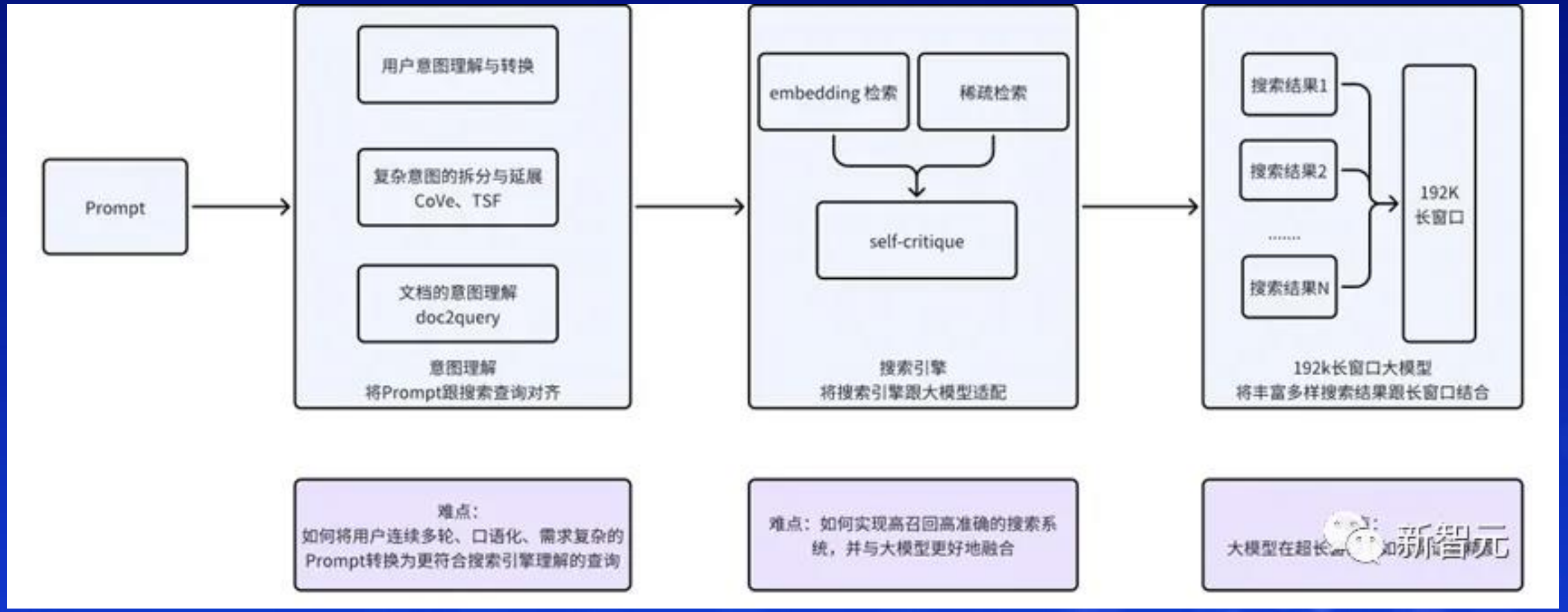


Advanced RAG



Modular RAG

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)

工作空间

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

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

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

权限管理

MLOps

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）

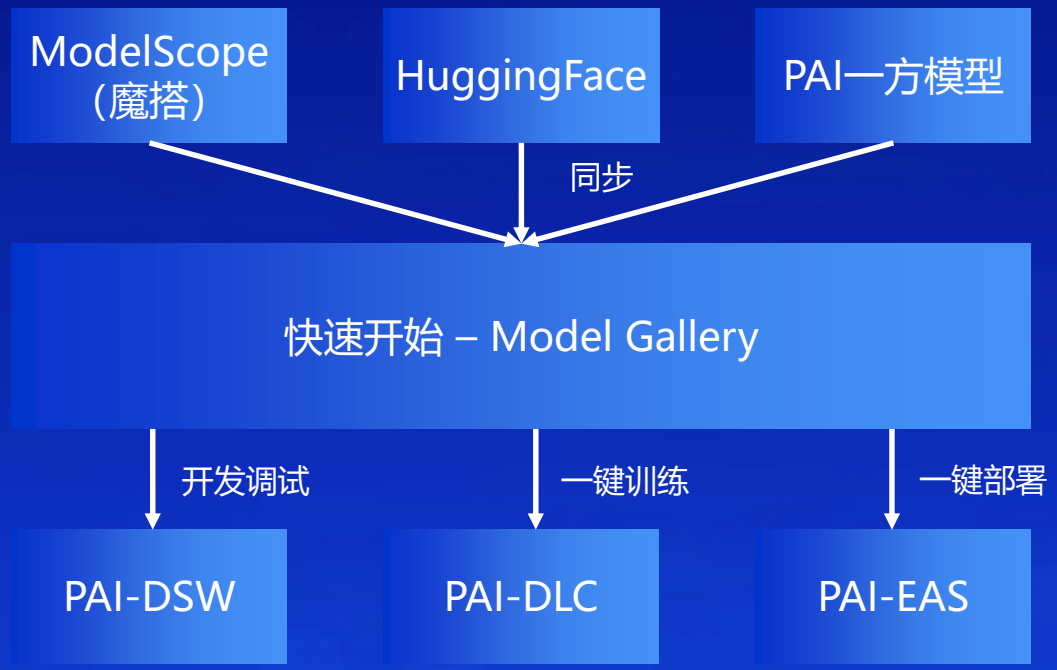
分布式存储 CPFS/NAS/OSS

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

QuickStart – 模型评测

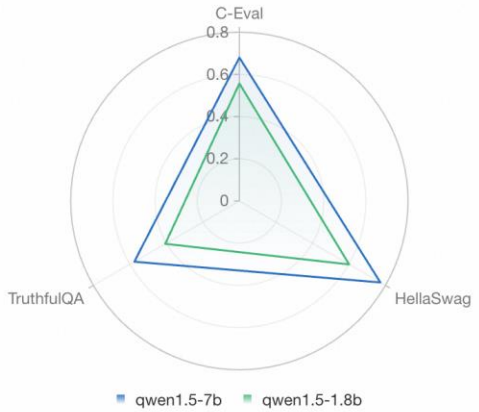
人工智能平台PAI / 模型评测 / 评测结果对比

← 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

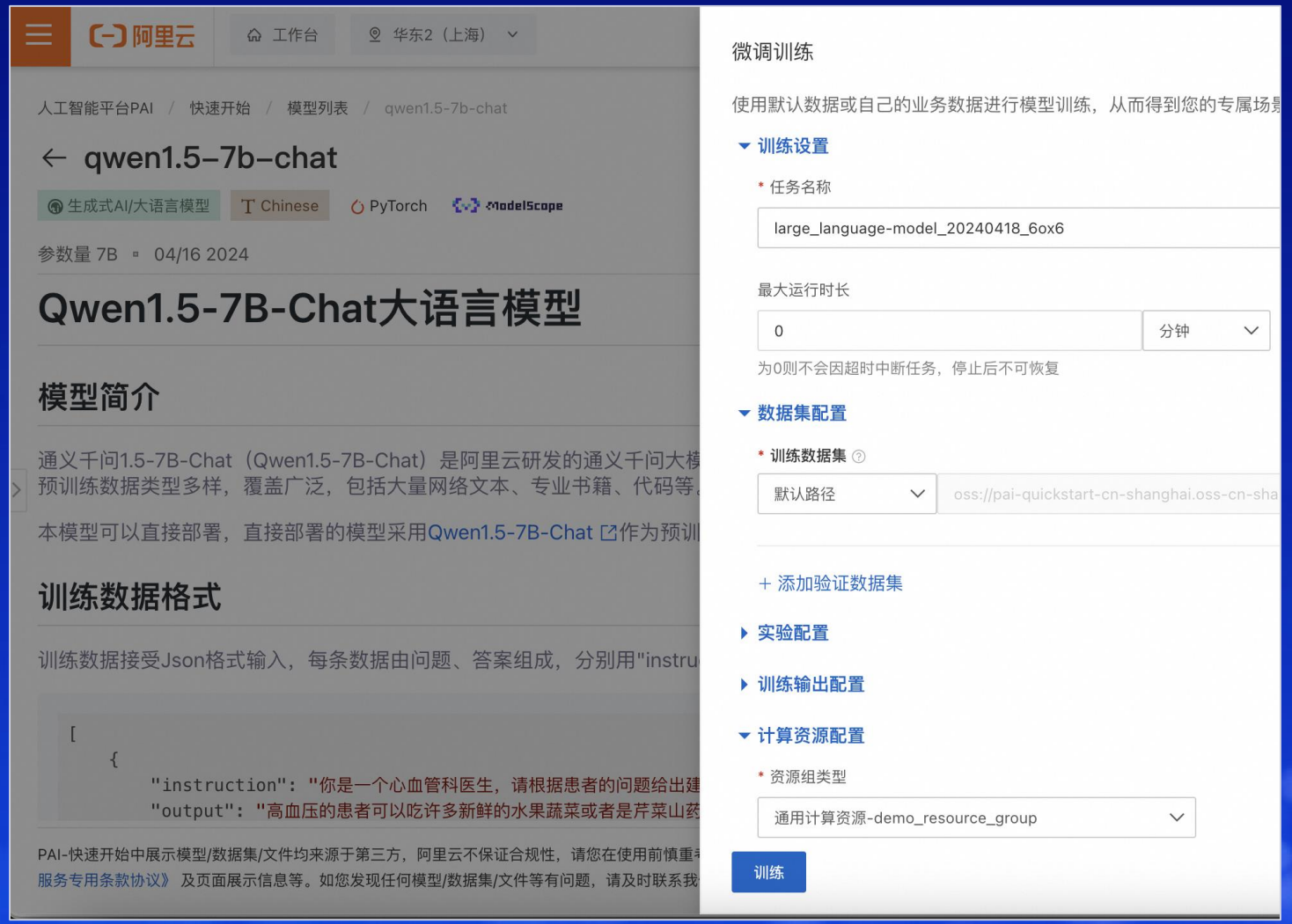
各数据集得分情况



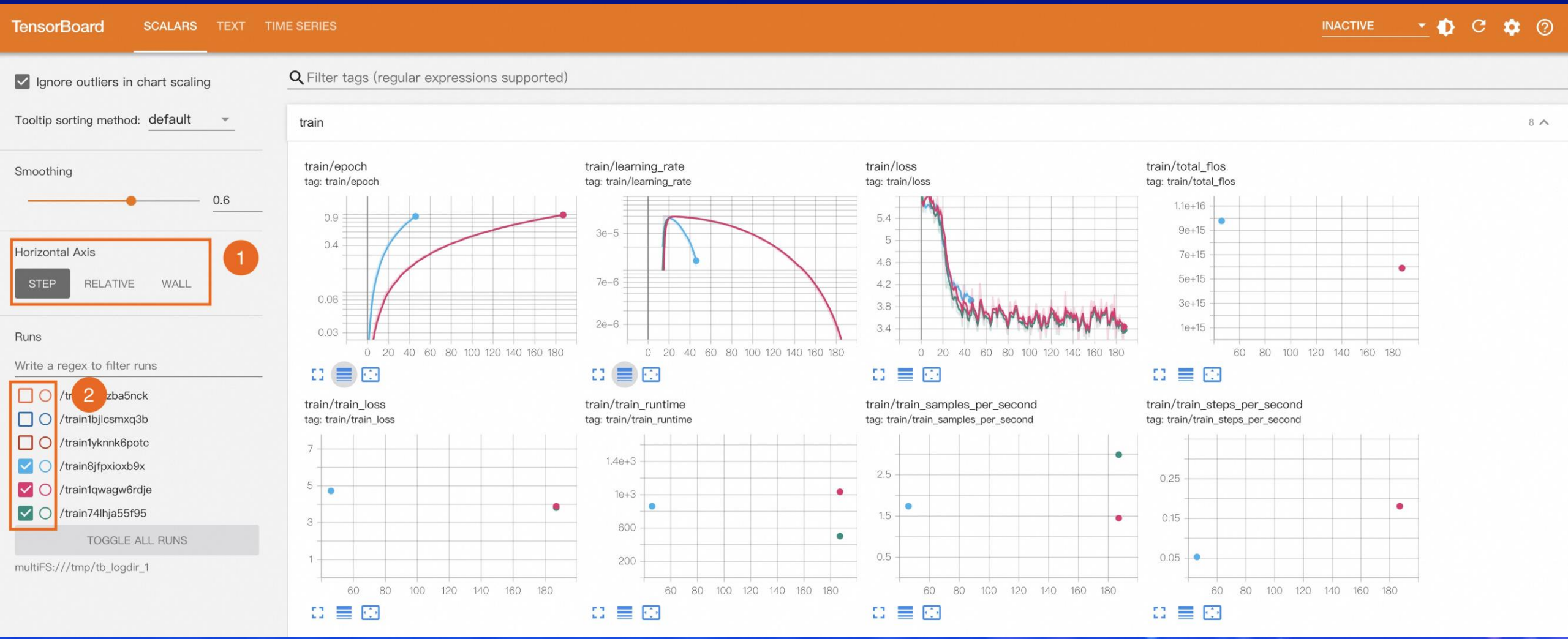
导出

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

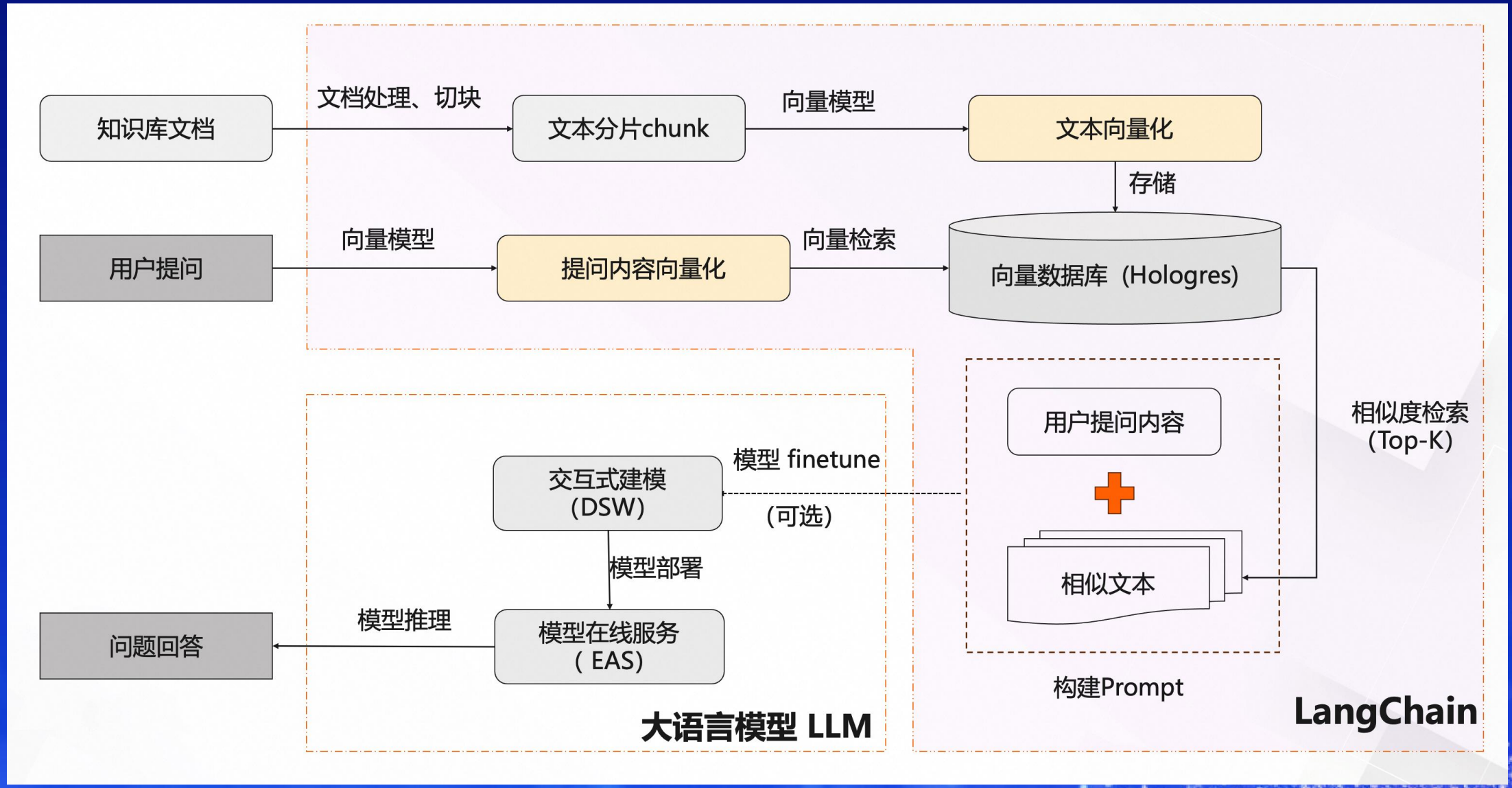
QuickStart – 模型微调



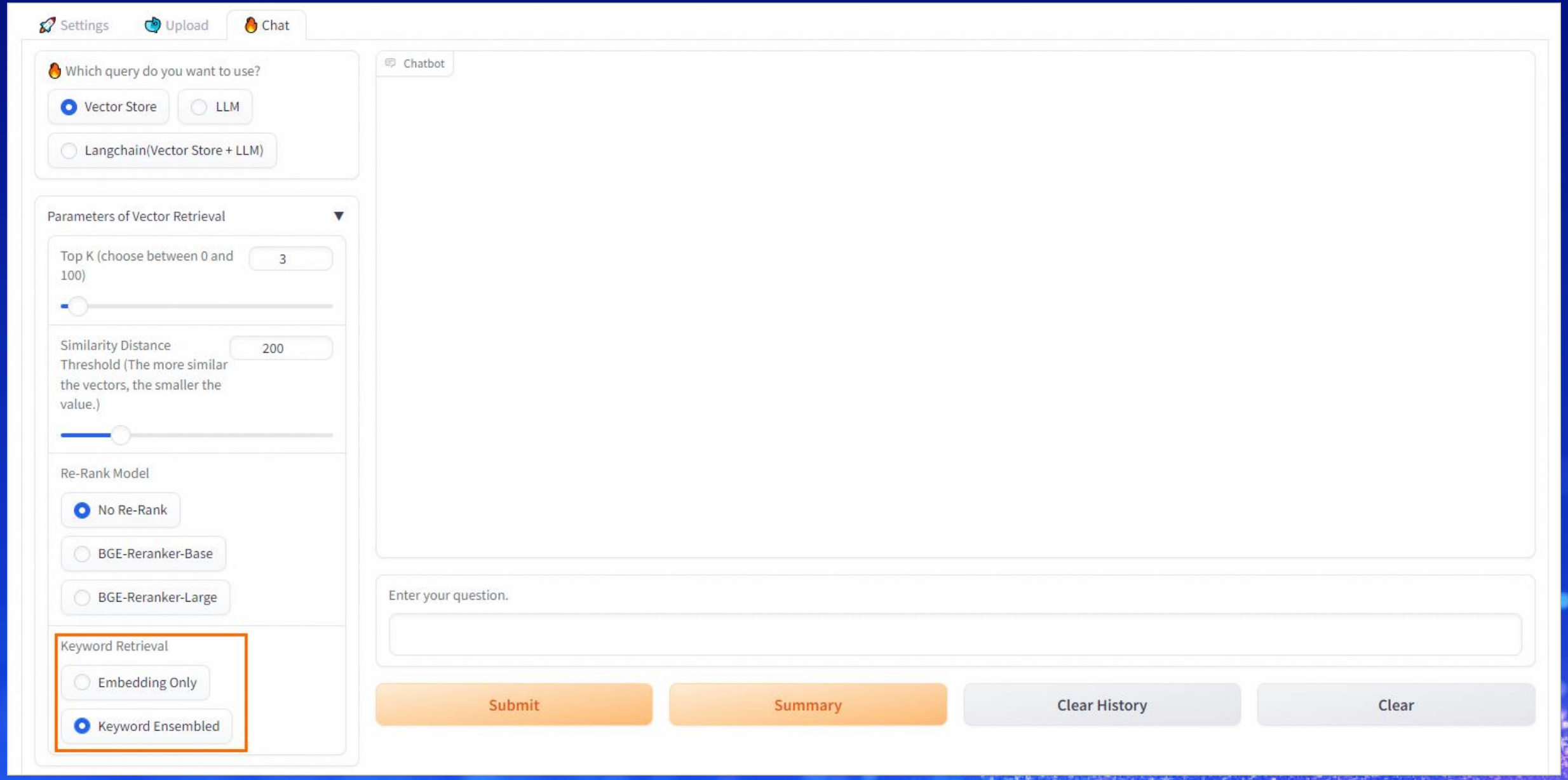
QuickStart – 实验管理



PAI RAG最佳实践


知识库文档
文档处理、切块
文本分片chunk
向量模型
文本向量化
存储
用户提问
向量模型
提问内容向量化
向量检索
向量数据库（Hologres）
用户提问内容
相似度检索
(Top-K)
交互式建模
(DSW)
模型 finetune
(可选)
模型部署
相似文本
问题回答
模型推理
模型在线服务
(EAS)
构建Prompt
大语言模型 LLM
LangChain

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)
3
Similarity Distance Threshold (The more similar the vectors, the smaller the value.)
200
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

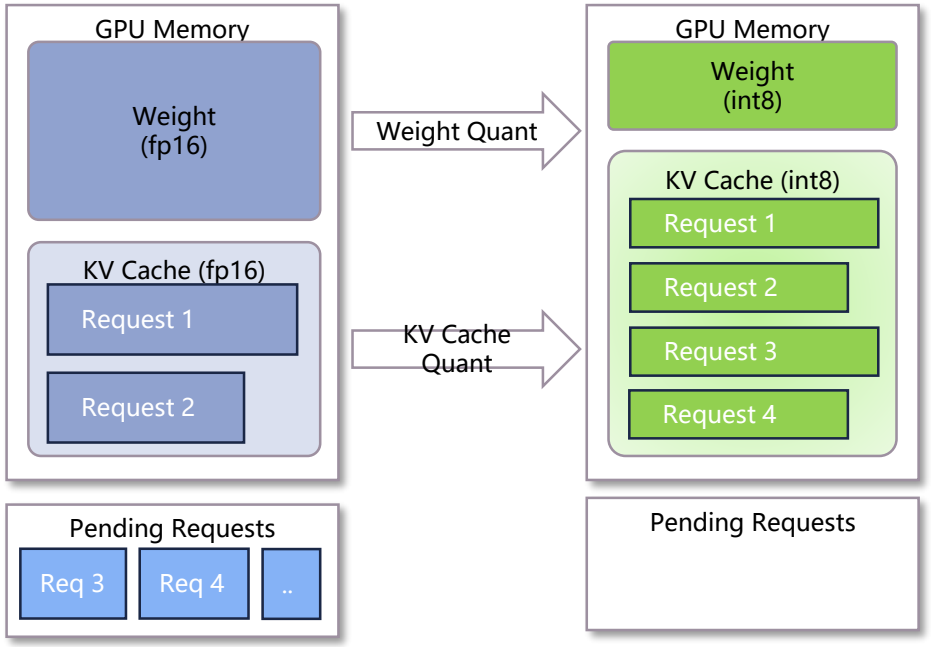
PAI LLM推理性能优化：BladeLLM

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

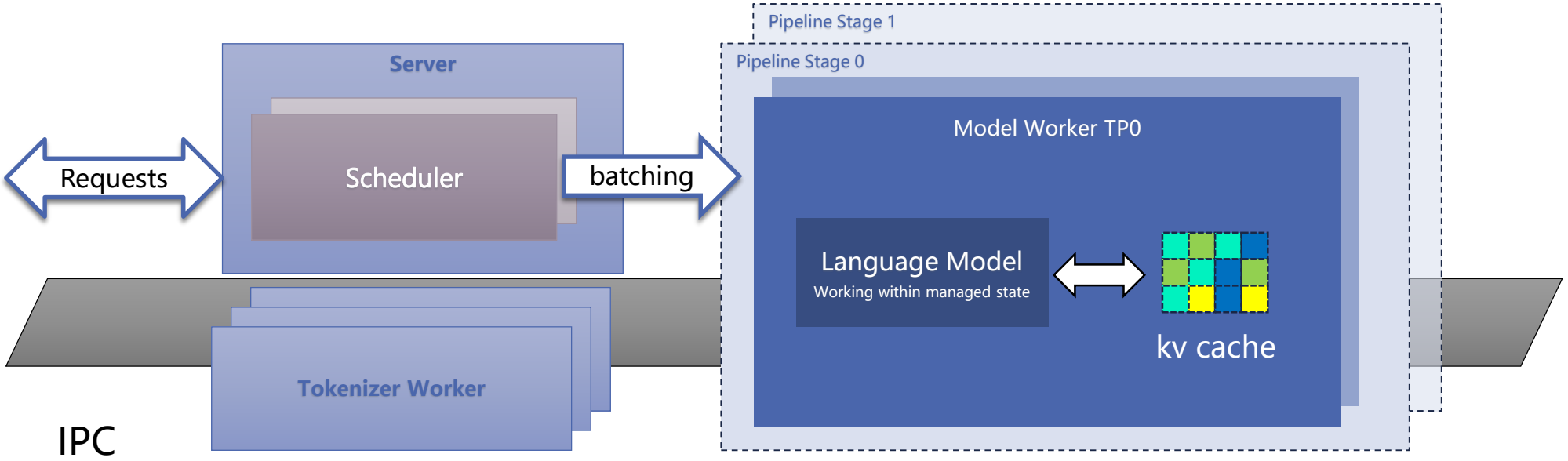
- 多样的量化算法策略和高效的量化算子，以适应调优需求提升精度和性能。



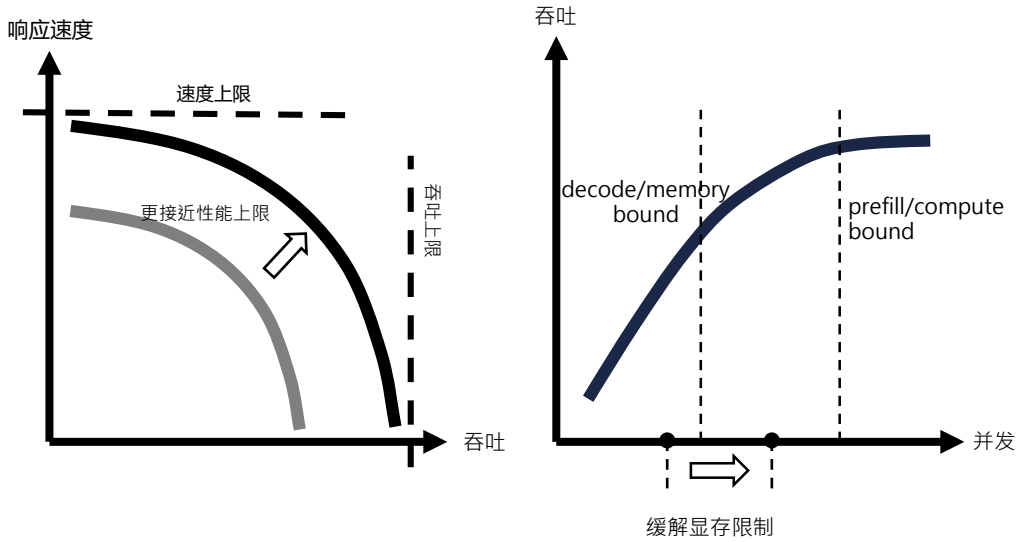
Quantization Mode	DType	Granularity	Mixed Precision
Weight-Only	int8/int4	per-channel/sub-channel	y
Activation-Weight	fp8/int8	per-channel/sub-channel	y
KV Cache	fp8/int8/int4	per-head/sub-head	y

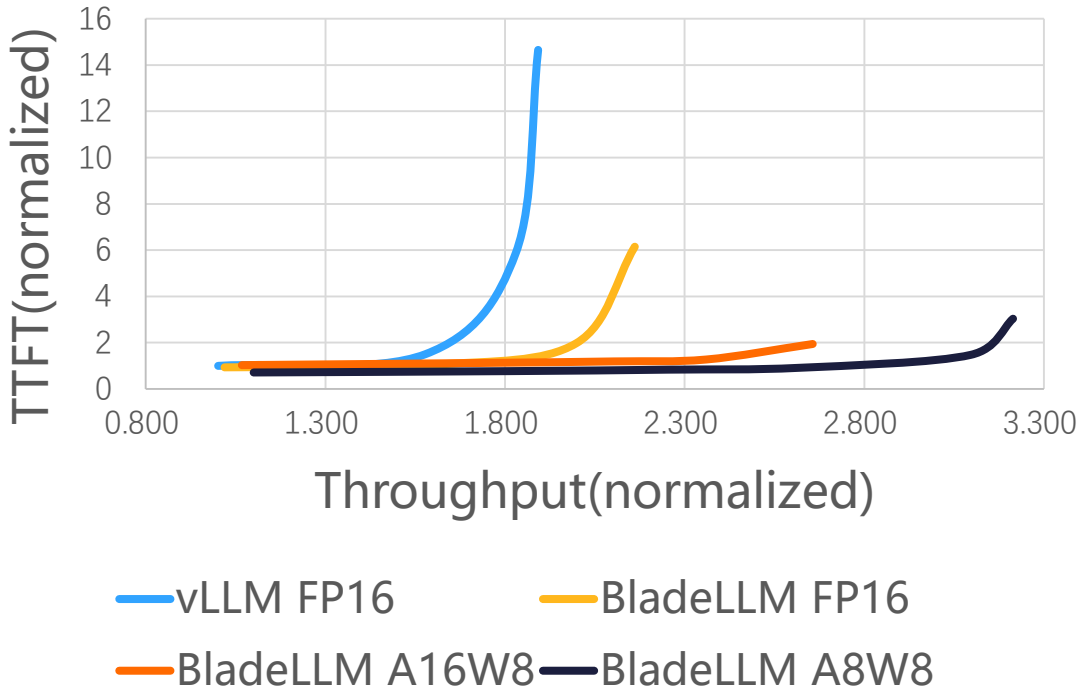
基础性能优化：高效进行时

- 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

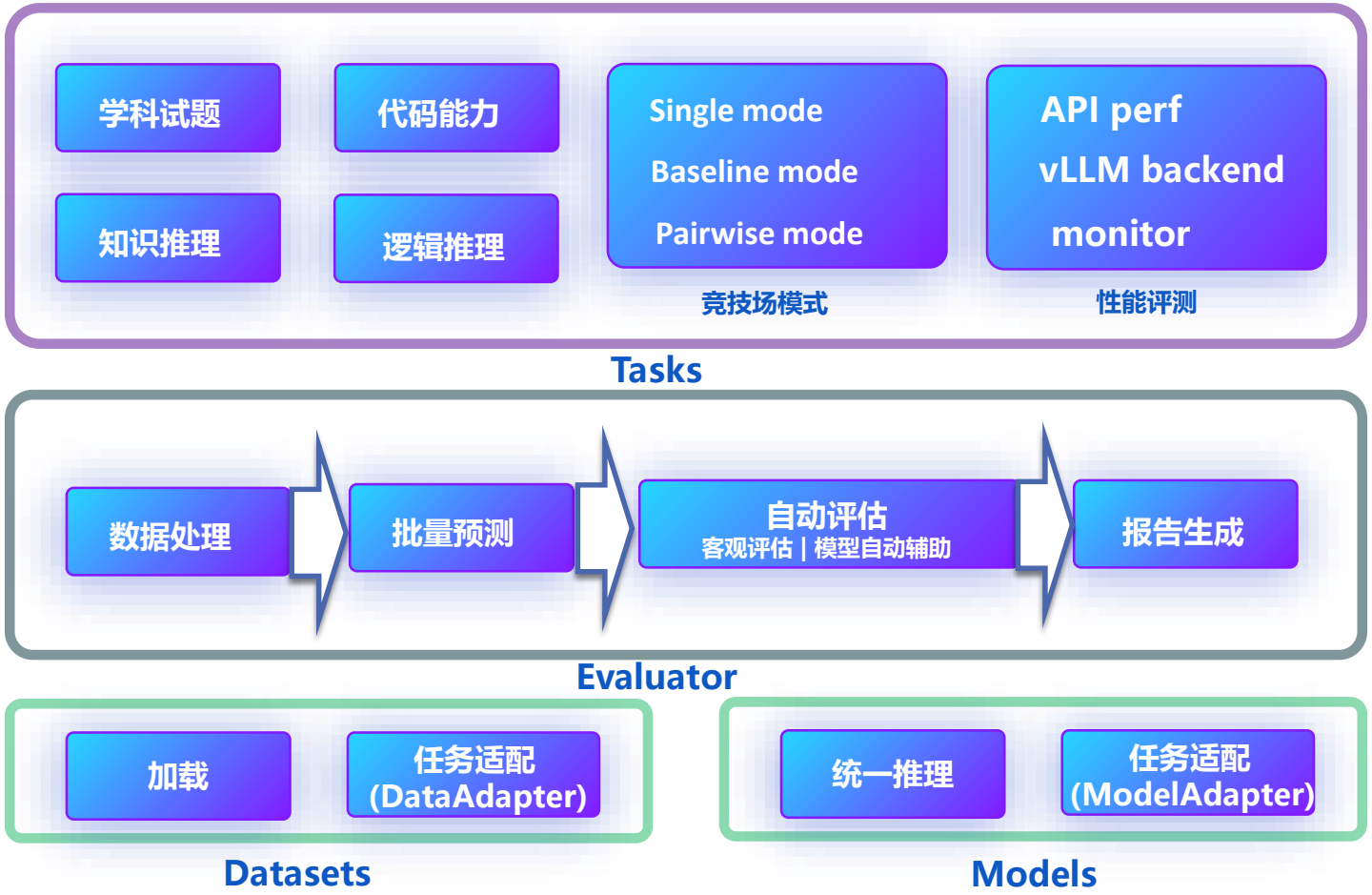




- 通过基础优化接近性能上限和缓解显存瓶颈
- 实际场景的性能提升示例

LLM评测框架：Eval-Scope

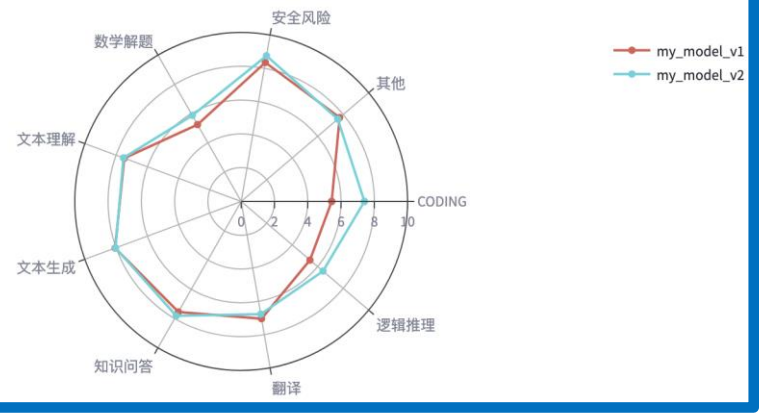
Eval-Scope是PAI与ModelScope共建的一款开源的轻量化LLM评测工具



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 + 技术团队的共同选择

KEYLINK ing

K+峰会

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

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