

大模型系列

内容介绍



ZOMI



AI 系统全栈架构



大模型 + AI系统全栈架构



大模型 + AI 系统技术栈

应用算法 (Algorithm)	领域算法	LLaMA、GPT、GLM、BLOOM、Reinforce Learning				
开发编程 (Development)	集成开发环境	编程环境: VS Code, Jupyter Notebook, PyCharm				
	编程语言	与主流语言集成: PyTorch and Tensorflow etc. inside Python				
使能 (Enable)	分布式框架	张量并行、数据并行、流水并行	张量数据库	RLHF	Prompt/Instruct Engineer	
框架 (Frameworks)	AI框架、推理引擎	中间表达 (IR) 、前端 API	编译优化 (Compilation)	图优化与图执行调度	推理优化	
		基本数据结构: 张量 (Tensor)	词法、语法、语义分析	计算图动静统一	常量折叠&冗余节点消除等	
		基本计算单元: 有向无环图 (DAG)	自动微分	计算图控制流	模型压缩: 蒸馏剪枝与量化	
异构与编译 (Compiler)	编程编译	前端优化	数据格式布局转换	内存分配	公共表达式消除	
		后端优化	代码优化、代码生成	算子循环优化	令和内存优化	
		编程模型: CUDA/Ascend C/Band C		AI编译器: TVM、XLA、TC		
		LLVM/GCC				
	底层通信	HCCL/NCCL、通信原语、集合通信算法				
体系结构 (Architecture)	计算节点	计算集群资源管理、作业调度、多级存储、千卡/万卡组网				
	底层硬件	AI芯片: CPU/GPU/AISC/FPGA/TPU/NPU		网络加速器: RDMA/IB/ROCK/NVLink/HCCS		

Course [chenzomil2.github.io](https://github.com/chenzomil2/github.io)



Talk Overview

1. AI 集群建设 : 计算、通信、存储的建设
2. 大模型数据 : 大模型数据集、数据处理、向量数据库
3. 大模型算法 : 从传统 NLP 到预训练 LLM 大模型
4. 大模型训练 : 大模型训练普通算法手段与稳定性分析
5. 分布式并行 : 模型并行、数据并行、优化器并行等
6. 大模型微调 : 全参微调、低参微调、指令微调算法
7. 大模型推理 : 量化压缩、长序列扩充推理、Cache 方法
8. 大模型评测 : NLP 下游任务、CV 下游任务、测评方案
9. 大模型智能体 : RLHF 流程、智能体、终身学习

Talk Overview

1. AI 集群建设：计算、通信、存储的建设
2. 大模型数据：大模型数据集、数据处理、向量数据库
3. 大模型算法：从传统 NLP 到预训练 LLM 大模型
4. 大模型训练：大模型训练普通算法手段与稳定性分析
5. 分布式并行：模型并行、数据并行、优化器并行等
6. 大模型微调：全参微调、低参微调、指令微调算法
7. 大模型推理：量化压缩、长序列扩充推理、Cache方法
8. 大模型评测：NLP 下游任务、CV 下游任务、测评方案
9. 大模型智能体：RLHF 流程、智能体、终身学习

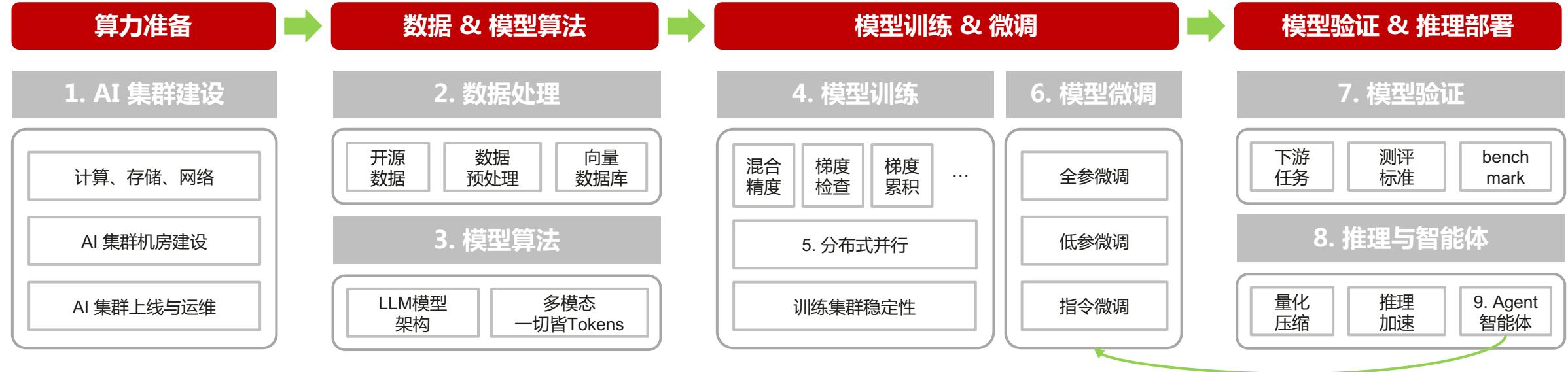
Talk Overview

- 1. AI 集群建设**：计算、通信、存储的建设
- 2. 大模型数据**：大模型数据集、数据处理、向量数据库
- 3. 大模型算法**：从传统 NLP 到预训练 LLM 大模型
- 4. 大模型训练**：大模型训练普通算法手段与稳定性分析
- 5. 分布式并行**：模型并行、数据并行、优化器并行等
- 6. 大模型微调**：全参微调、低参微调、指令微调算法
- 7. 大模型推理**：量化压缩、长序列扩充推理、Cache方法
- 8. 大模型评测**：NLP 下游任务、CV 下游任务、测评方案
- 9. 大模型智能体**：RLHF 流程、智能体、终身学习

大模型全流程



大模型业务全流程

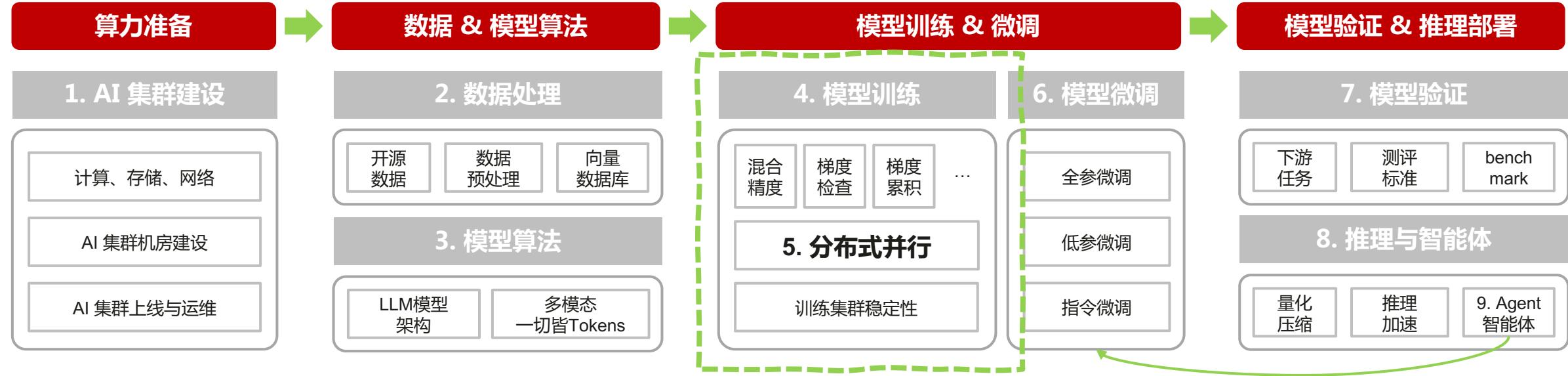


大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

5. 分布式并行



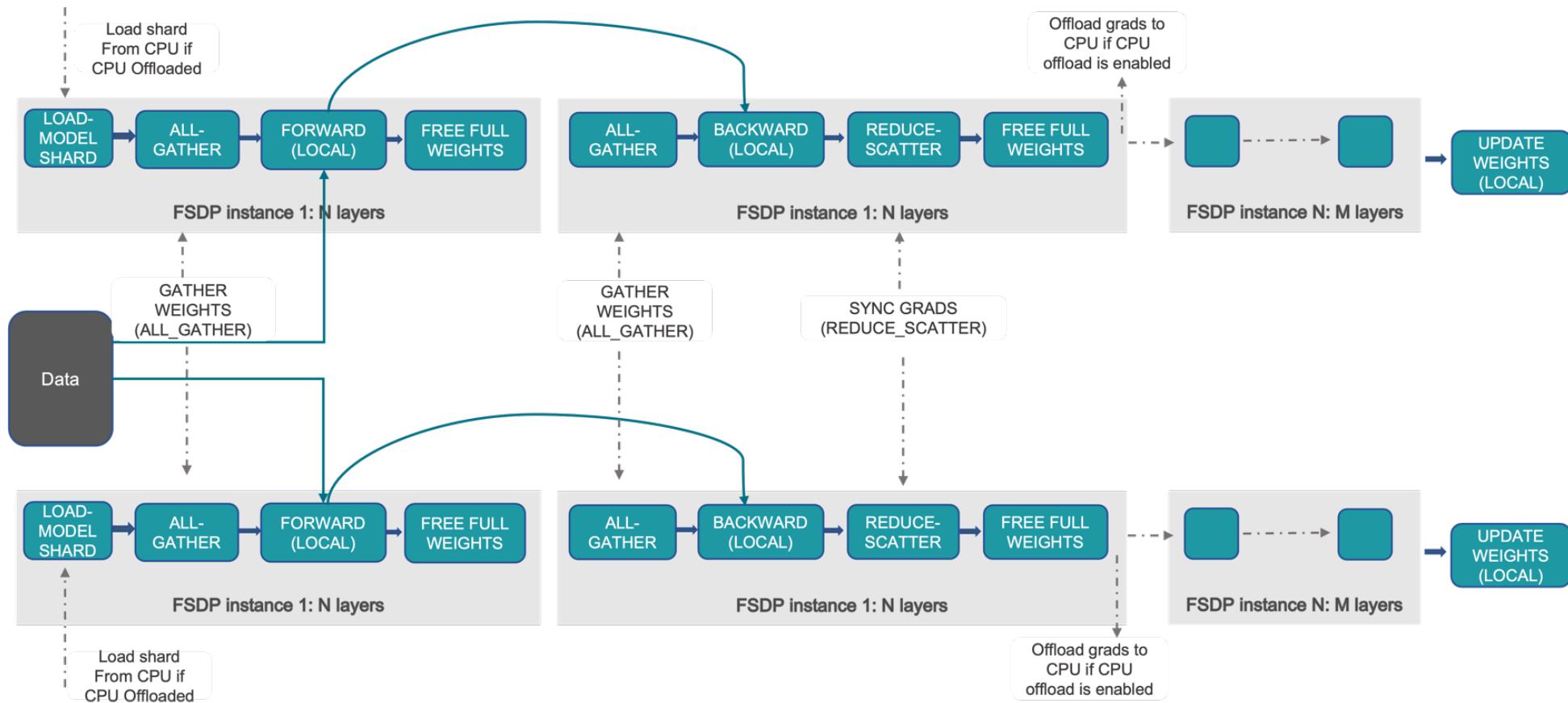
大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

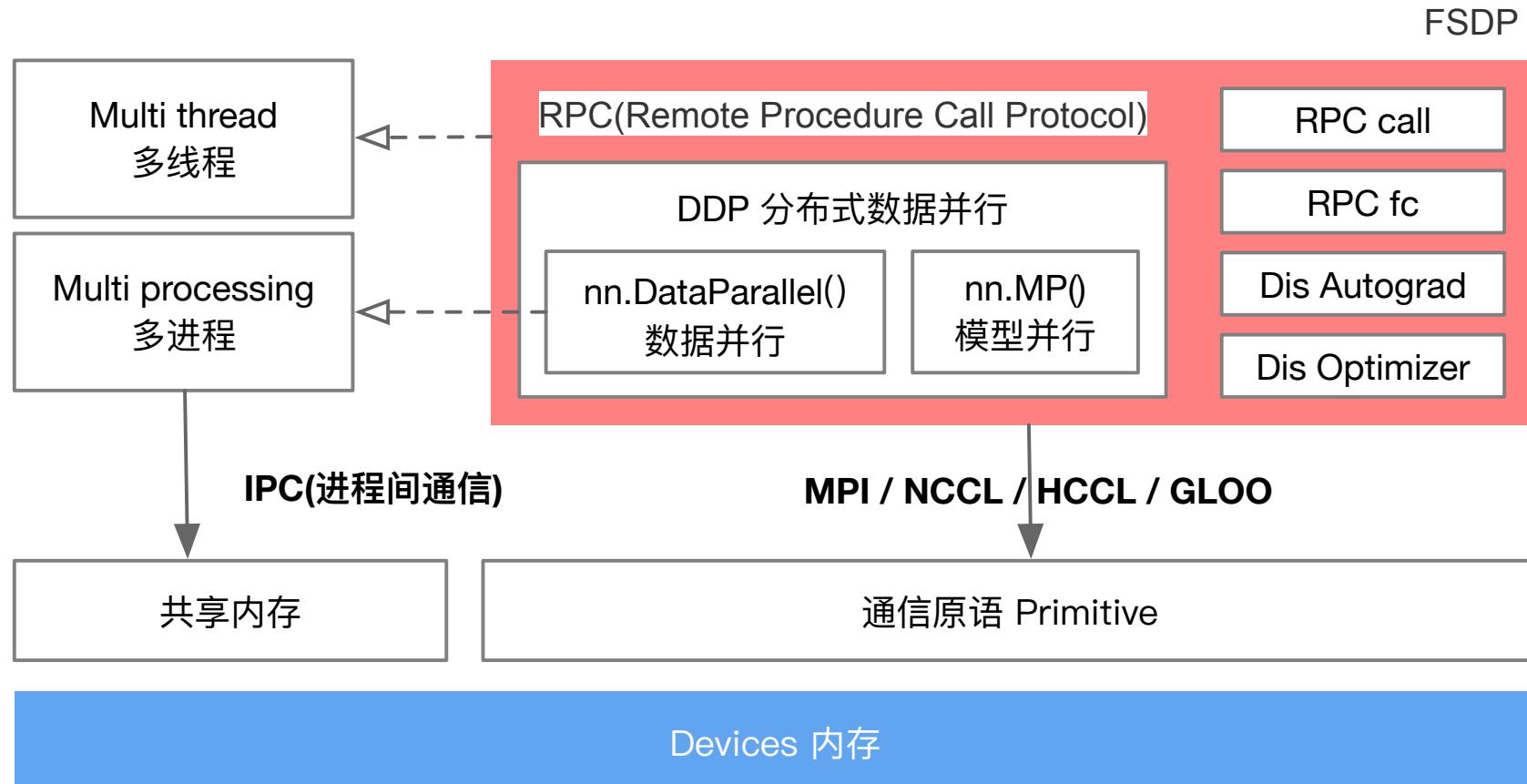
分布式并行

- 模型太大了放不下，如何进行分布式并行？模型和参数是怎么切分？有什么规律？



分布式并行

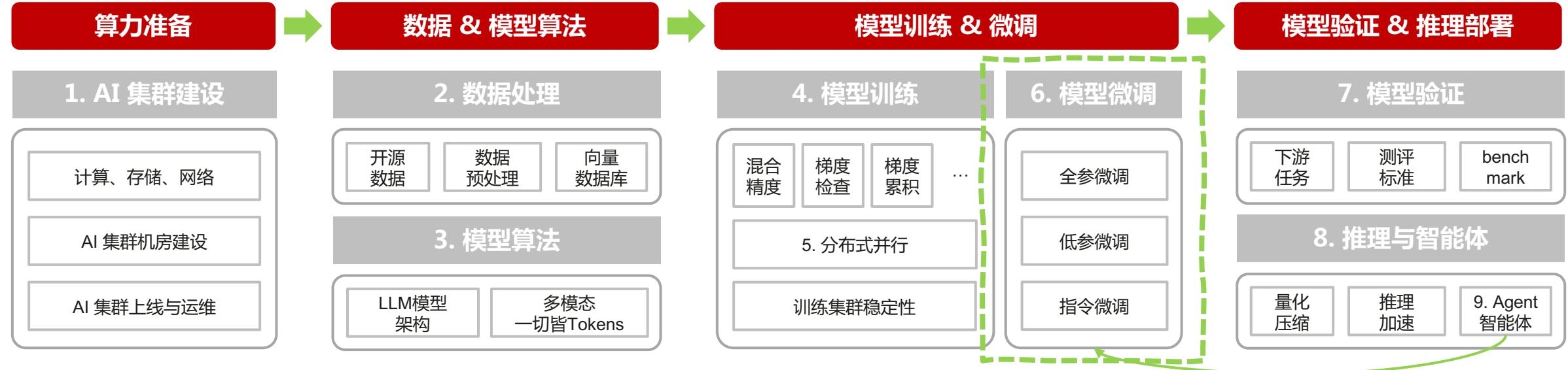
- 不能只关注于分布式并行算法，更要关注分布式框架整体提供的服务



6. 模型微调



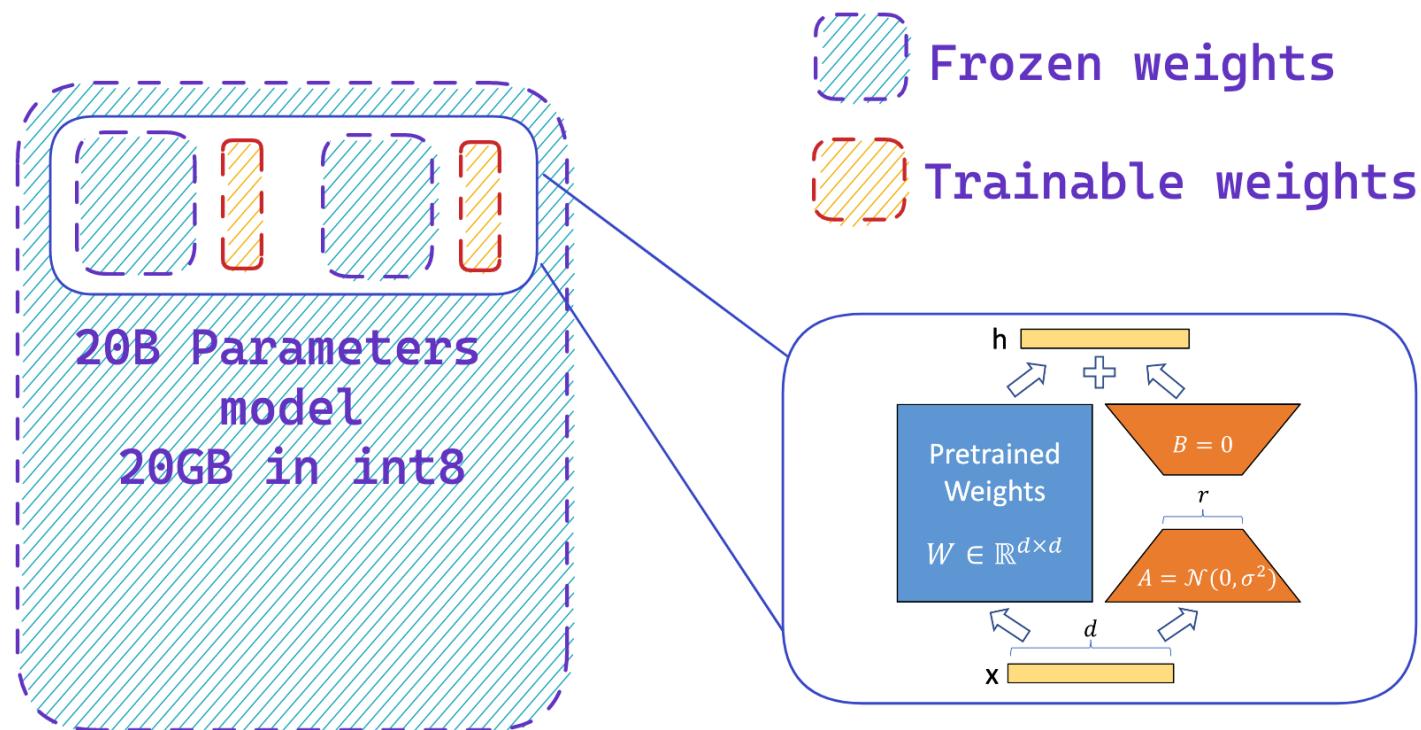
大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

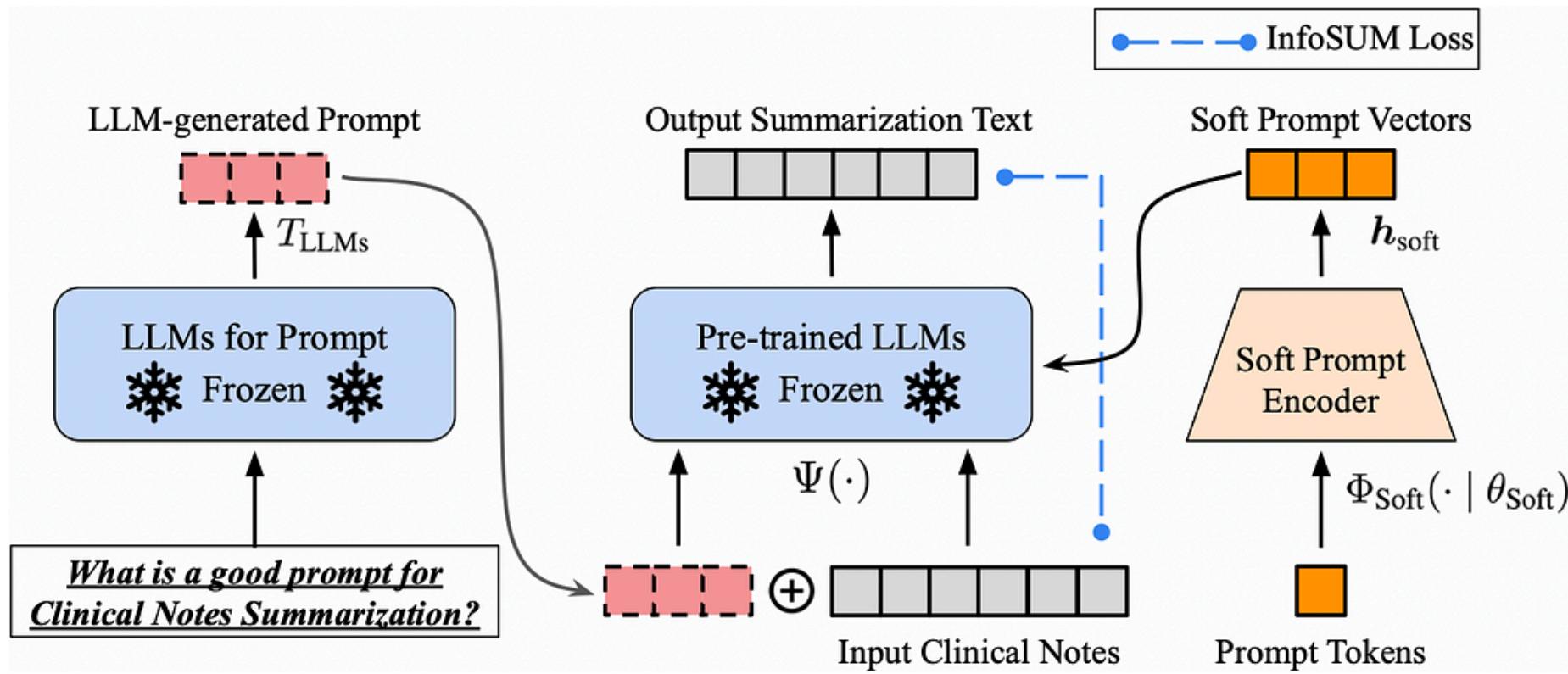
模型微调

- 全参微调跟预训练什么区别？低参微调有哪些典型算法？



模型微调

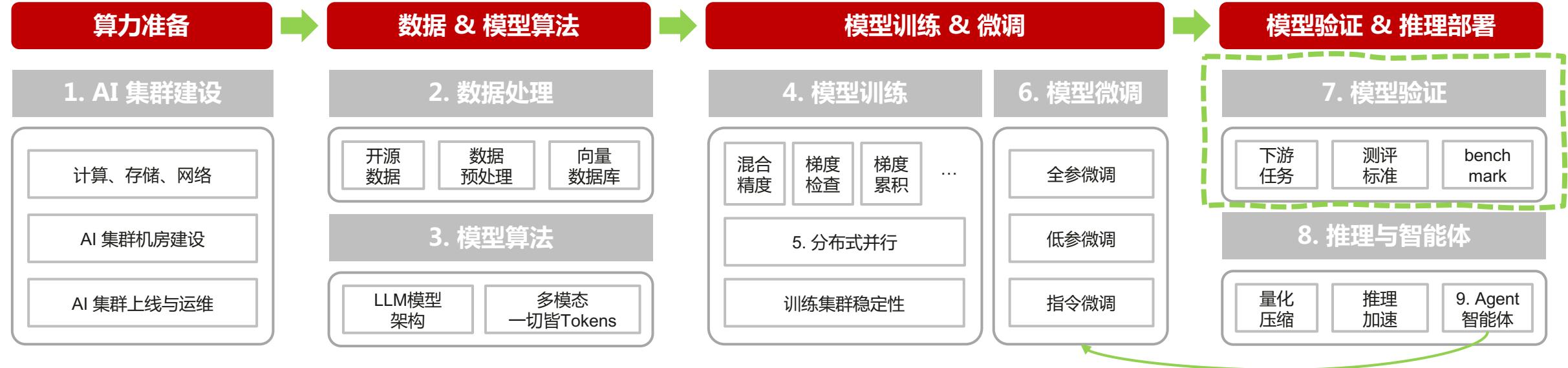
- 指令微调和 Prompt 微调怎么区分？对大模型的学习和微调有哪些作用？



7. 大模型验证



大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

评价体系

- 大模型缺乏统一的标准评价体系，各研究机构和高校争相推出自己的大模型 Benchmark 测评。

智库榜单

榜单/测评集名称	Owner
新华社研究院	
FlagOPEN	FlagAI、FlagData、FlagEval 北京智源研究院

学术榜单

榜单/测评集名称	Owner
MMLU	加州大学伯克利分校
C-Eval	清华大学、上海交通大学
CMMLU	上海交通大学、微软亚洲研究院
OpenCompass	上海人工智能实验室
MME	腾讯优图实验室、厦门大学
KOLA	清华大学
AgentBench	清华大学、加州大学伯克利分校
AplacaEval	斯坦福大学
GLUE	NLP领域权威榜单，纽约大学、DeepMind

企业榜单

榜单/测评集名称	Owner
AGIEval	
Big-Bench Hard	谷歌研究院
GSM8K	OpenAI
HuamEval	OpenAI

社区榜单

榜单/测评集名称	Owner
SuperCLUE	CLUE学术社区



评价体系

- 训练完一个大模型，LOSS有精度差异，如何系统性评价生成式模型的好坏？

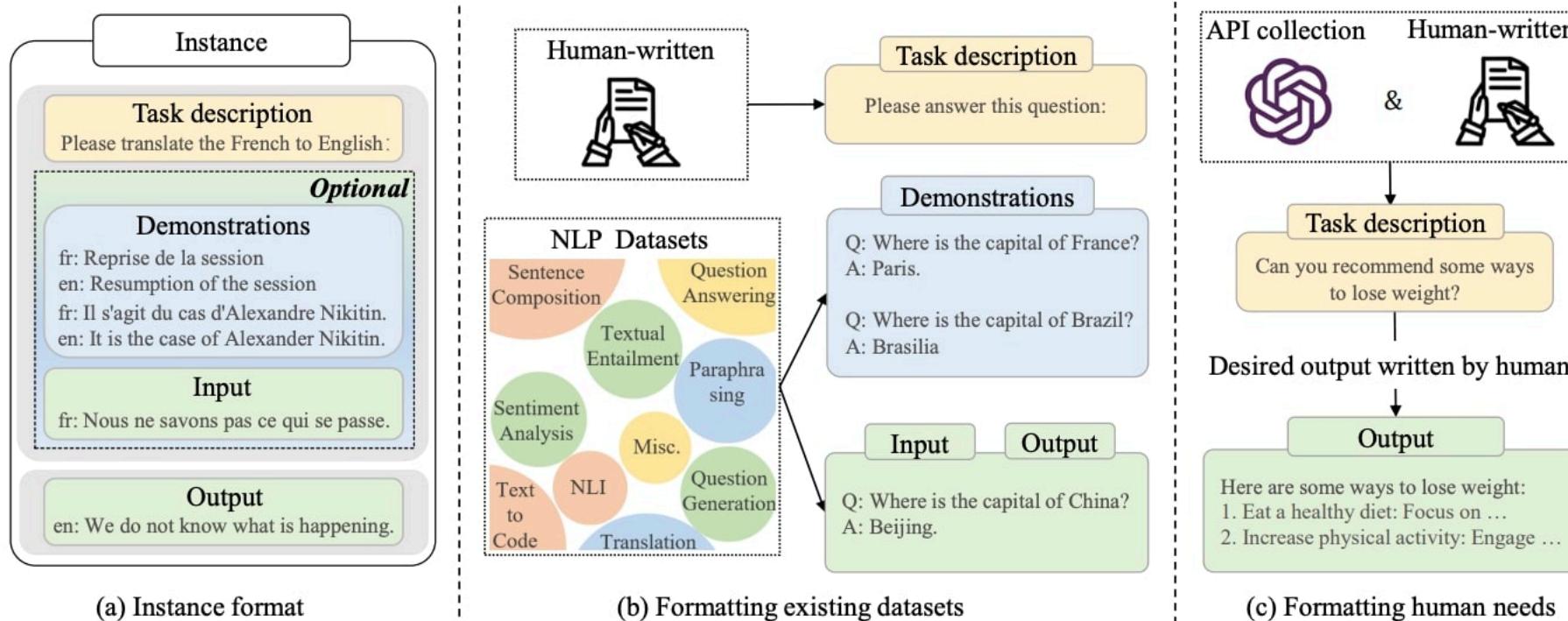


Fig. 4. An illustration of instance formatting and two different methods for constructing the instruction-formatted instances.

评价体系

- 训练完一个大模型，LOSS有精度差异，如何系统性评价生成式模型的精度情况？

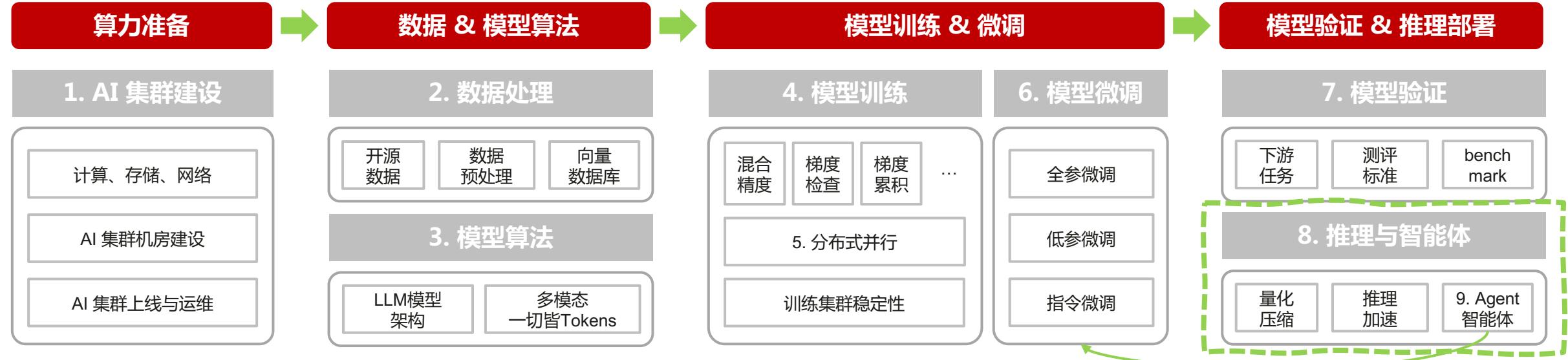
		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	88.0	82.3	-	83.4	81.1	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	80.0	57.8	58.6
	65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	60.2

Table 3: Zero-shot performance on Common Sense Reasoning tasks.

8. 大模型推理



大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

大模型推理

- 一次10万token ! GPT4最强对手史诗升级，百页资料一分钟总结完毕！
- 腾讯混元大模型今日起对外开放，每1000 token 收费0.14 元
- 真·量子速读：突破GPT-4一次只能理解50页文本限制，新研究扩展到百万token !

延迟
(Latency)

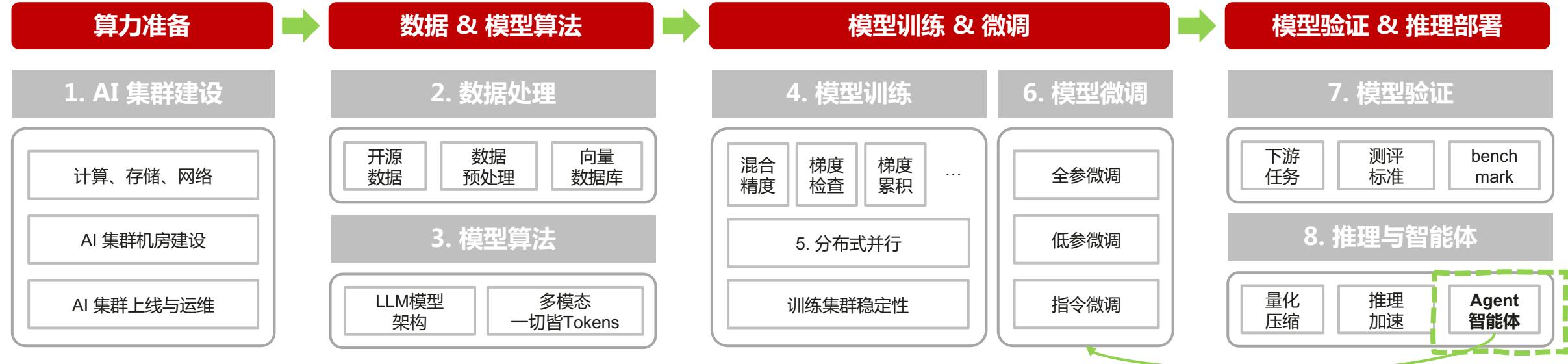
吞吐量
(Throughput)

QPS
(Queries Per Second)

9. Agent 智能体



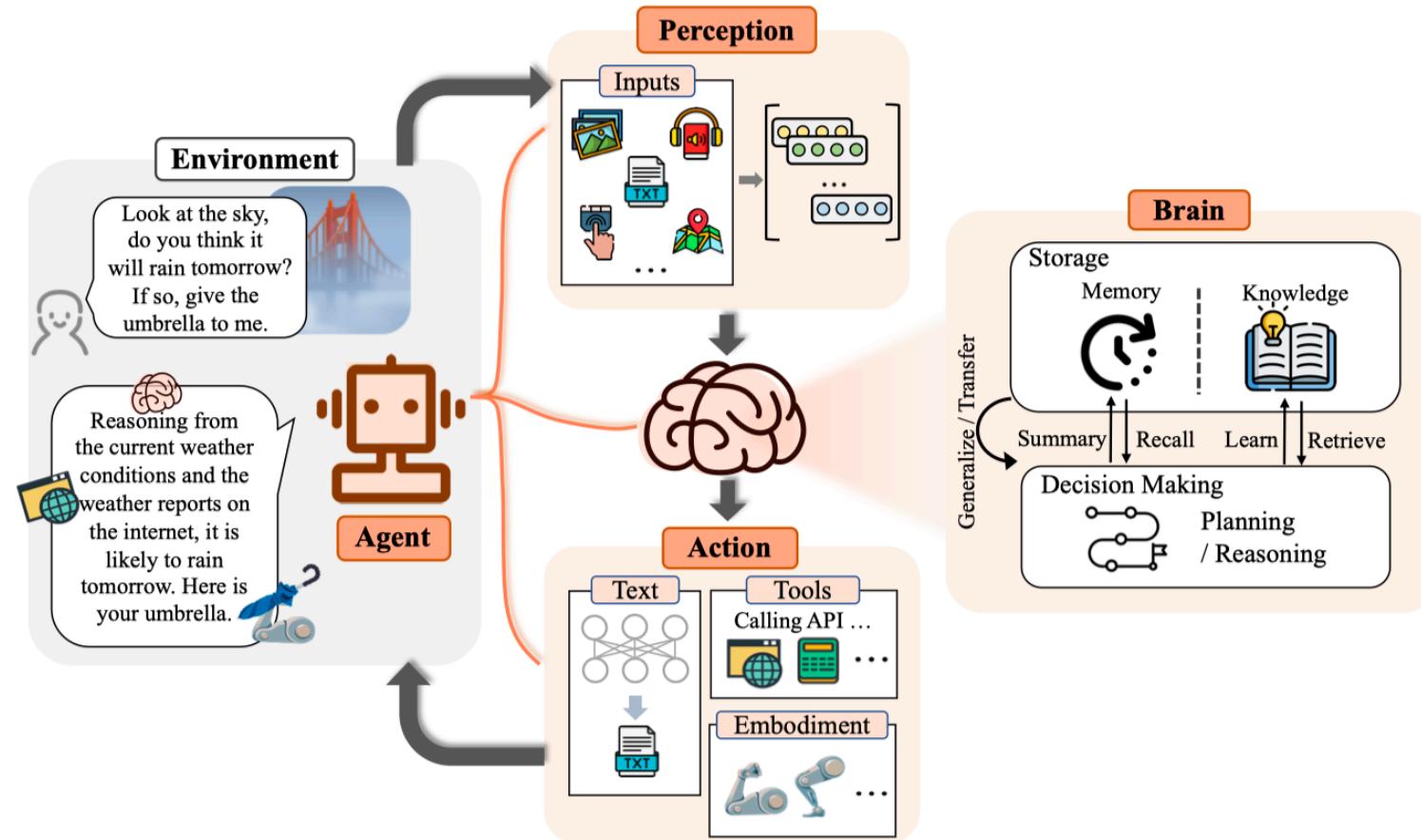
大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

Agent 智能体

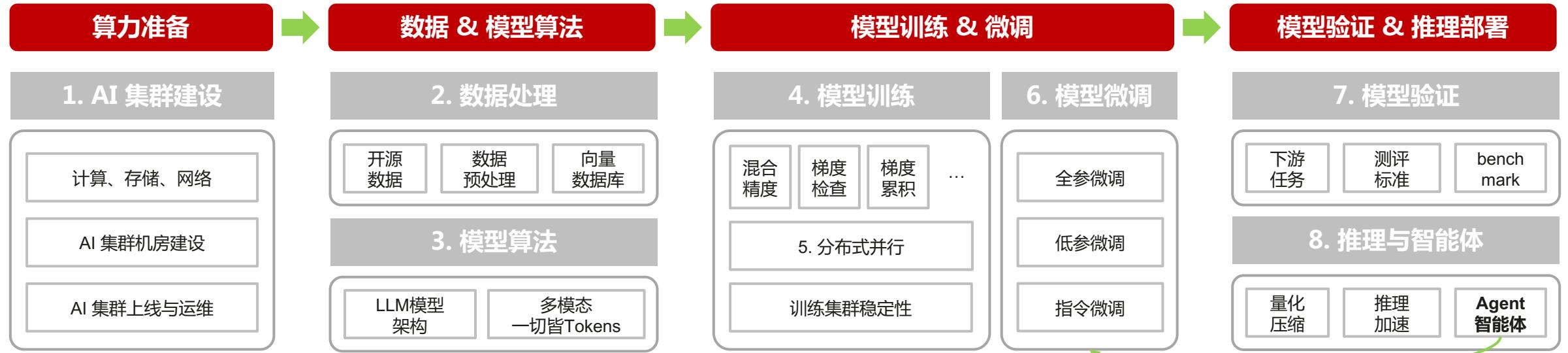
- GAI是持续学习、终身学习，大模型最终形态给GAI提供基座，实现智能体自学习需要哪些技术？



总结



Talk Review



1. **AI 集群建设** : 计算、通信、存储的建设
2. **大模型数据** : 大模型数据集、数据处理、向量数据库
3. **大模型算法** : 从传统 NLP 到预训练 LLM 大模型
4. **大模型训练** : 大模型训练普通算法手段与稳定性分析

5. **分布式并行** : 模型并行、数据并行、优化器并行等
6. **大模型微调** : 全参微调、低参微调、指令微调算法
7. **大模型推理** : 量化压缩、长序列扩充推理、Cache方法
8. **大模型评测** : NLP 下游任务、CV 下游任务、测评方案
9. **大模型智能体** : RLHF 流程、智能体、终身学习



Thank you

把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem