大模型
内容介绍

ZOMI

Course chenzomi12.github.io

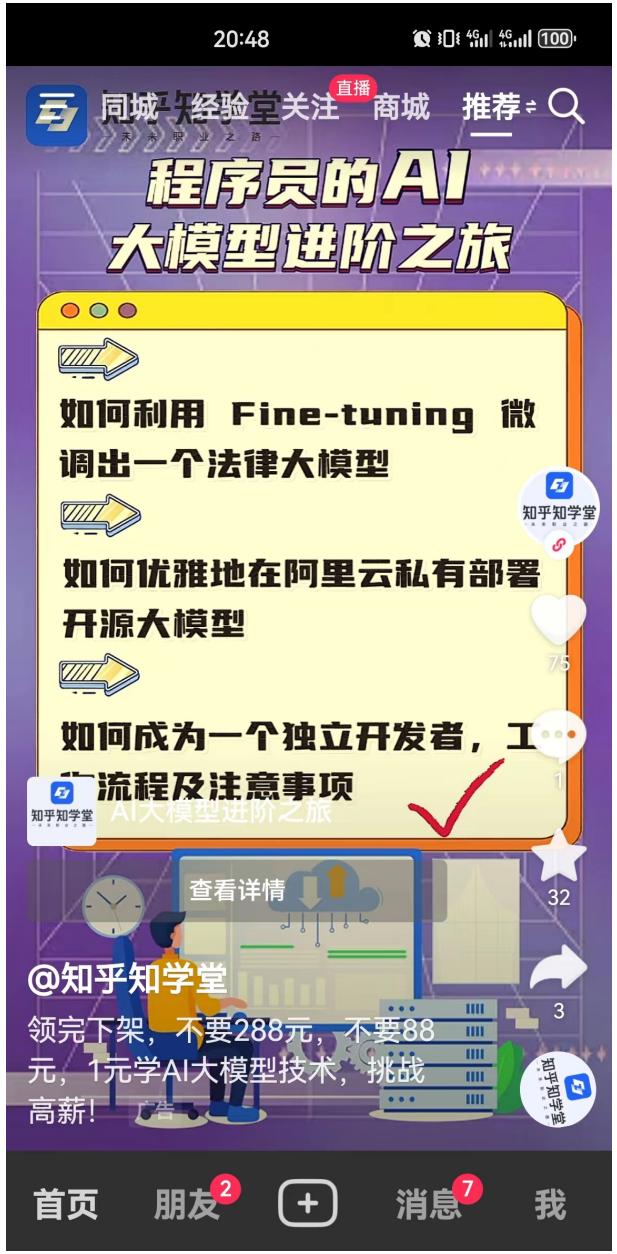
GitHub github.com/chenzomi12/DeepLearningSystem

AI 系统全栈架构


AI系统全栈架构图
应用
CV
NLP
Audio
...
开发
编程语言（AI与主流语言集成）
python
框架
PyTorch
TensorFLow
Caffe
MindSpore
JAX
MNN
ONNX
TensorRT
AI框架
AI推理引擎
编译
编程
TVM
GLOW
XLA
TC
CUDA
LLVM
GCC
AI编译器
传统编译器
AI芯片
GPU
TPU
IPU
NPU
NPU
CPU


21:54
同城 经验 关注 商城 推荐
坚持学习2天
AI大模型全栈工程师
你会感谢你自己
10
1
4
1
@知乎知学堂
领完下架，不要288元，不要88元，1元学AI大模型技术，挑战高薪！ 广告
前置学习条件：本科及以上学历，有编程经验即可
首页 朋友 1 消息 6 我


12:52
同城 经验 关注 商城 推荐
工作两年的程序员
学习AI大模型技术三个月
努力实现技术致富
28
5
8
2
@知乎知学堂
领完下架，不要288元，不要88元，1元学AI大模型技术，挑战高薪！ 广告
前置学习条件：本科及以上学历，有编程经验即可
首页 朋友 红包 消息 3 我


20:48
同城 经验 关注 直播 商城 推荐
程序员的AI
大模型进阶之旅
如何利用 Fine-tuning 微调出一个法律大模型
如何优雅地在阿里云私有部署开源大模型
如何成为一个独立开发者，工
流程及注意事项
查看详情
75
1
32
3
@知乎知学堂
领完下架，不要288元，不要88元，1元学AI大模型技术，挑战高薪！ 广告
首页 朋友 2 消息 7 我

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	

Talk Overview

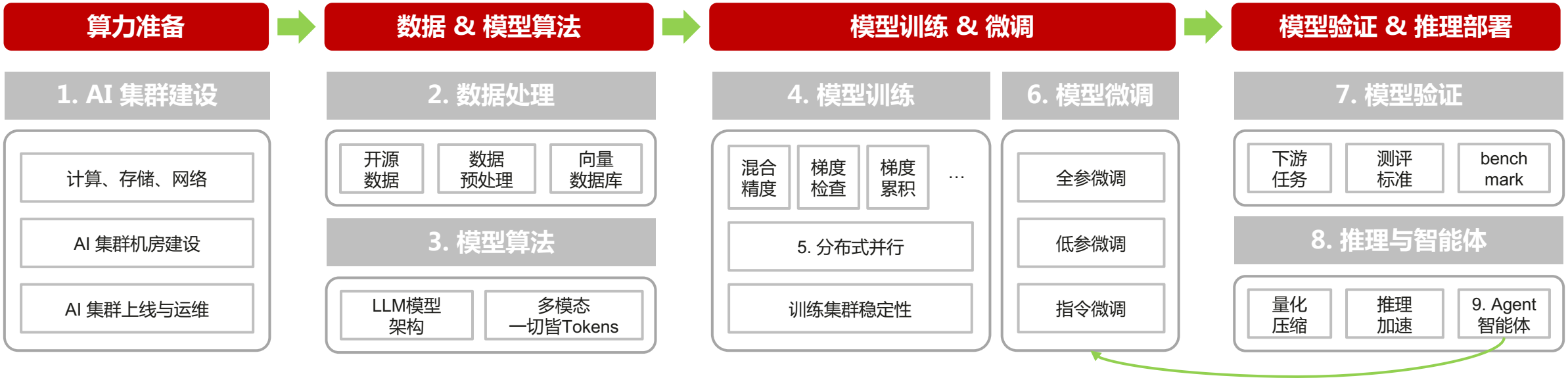
1. **AI 集群建设**：计算、通信、存储的建设
2. **大模型数据**：大模型数据集、数据处理、向量数据库
3. **大模型算法**：从传统 NLP 到预训练 LLM 大模型
4. **大模型训练**：大模型训练普通算法手段与稳定性分析
5. **分布式并行**：模型并行、数据并行、优化器并行等
6. **大模型微调**：全参微调、低参微调、指令微调算法
7. **大模型推理**：量化压缩、长序列扩充推理、Cache方法
8. **大模型评测**：NLP 下游任务、CV 下游任务、测评方案
9. **大模型智能体**：RLHF 流程、智能体、终身学习

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

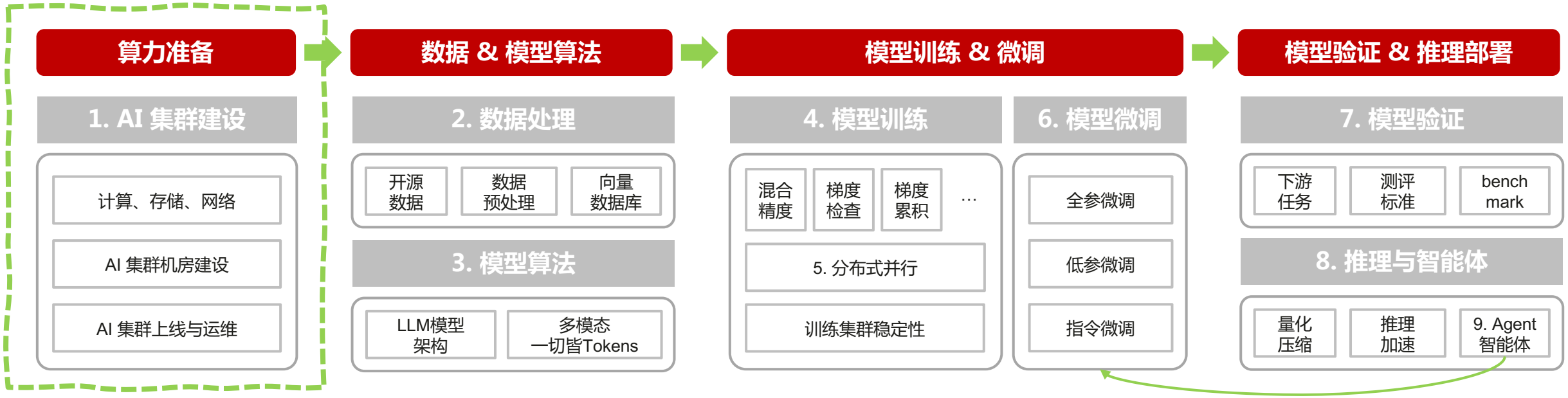


大模型不仅需要 LLM 算法，同时需要提供

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

1. AI 集群建设

大模型业务全流程

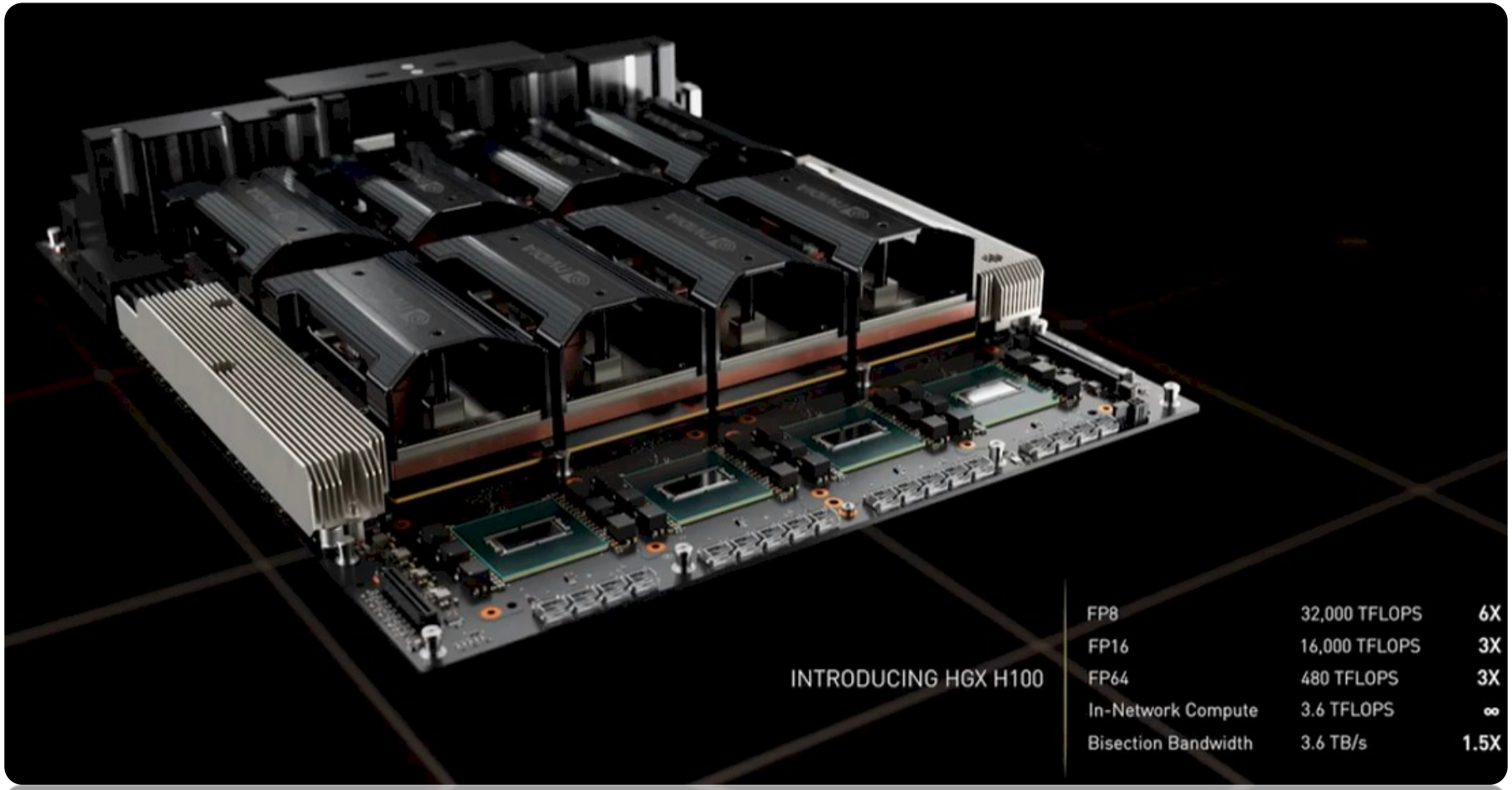


大模型不仅需要 LLM 算法，同时需要提供

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

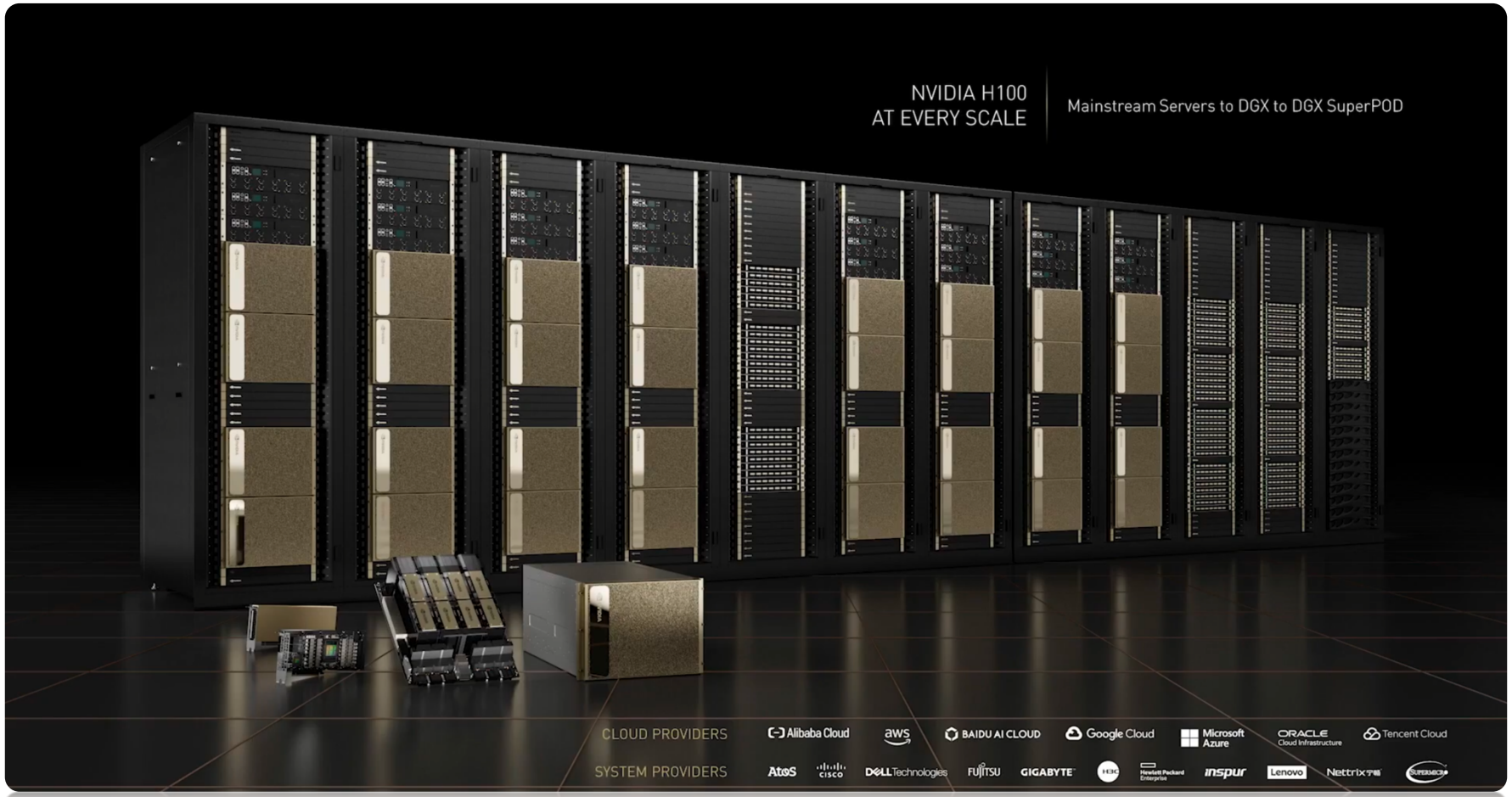
1. AI 集群建设

AI 芯片提供澎湃的算力，大模型对 AI 芯片提出哪些新的需求？



1. AI 集群建设

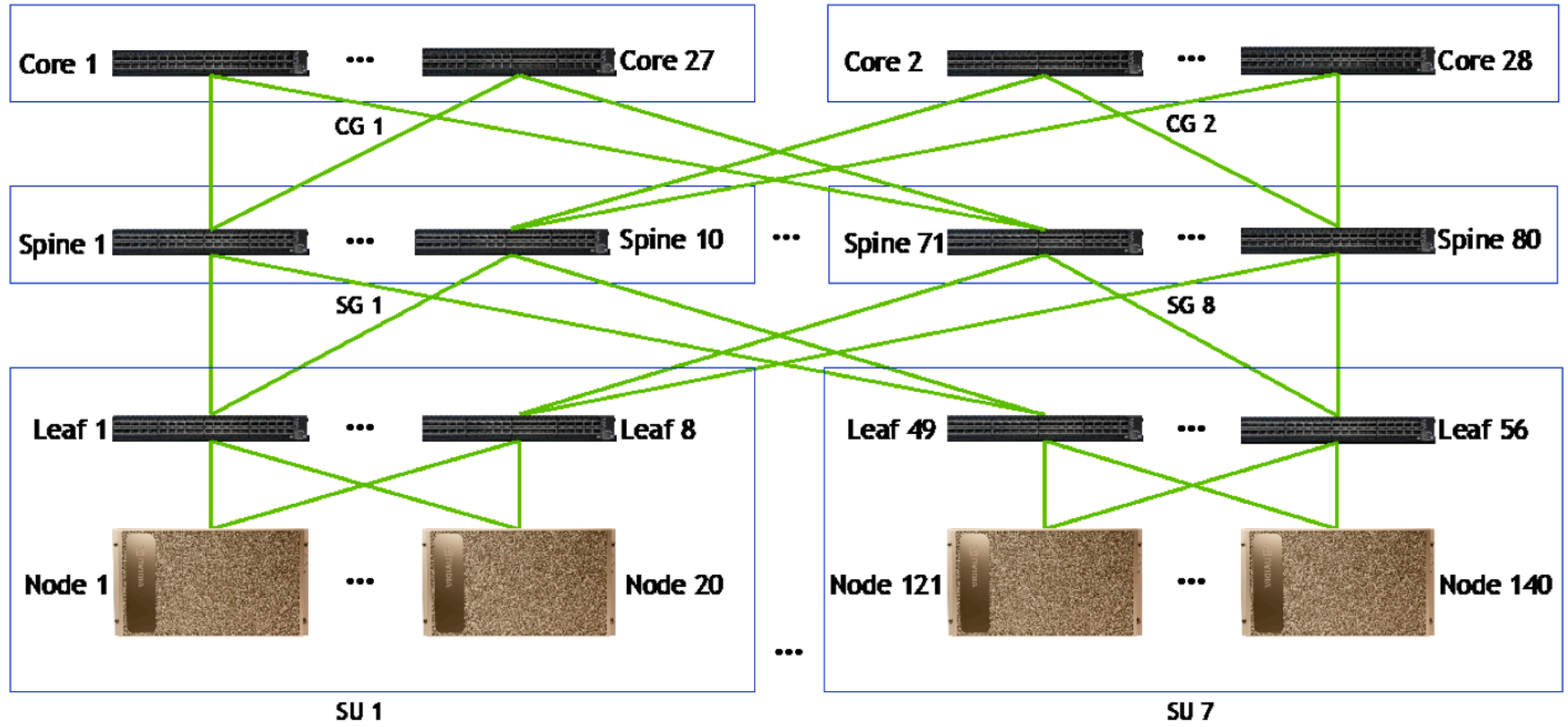
大模型涌现，以 AI 集群中心提供 E 级算力，那么 AI 集群建设跟传统的云中心建设什么区别？



大模型业务全流程：AI 集群建设

如何组网才能减少网络通信的消耗，提升大模型对算力的利用率？

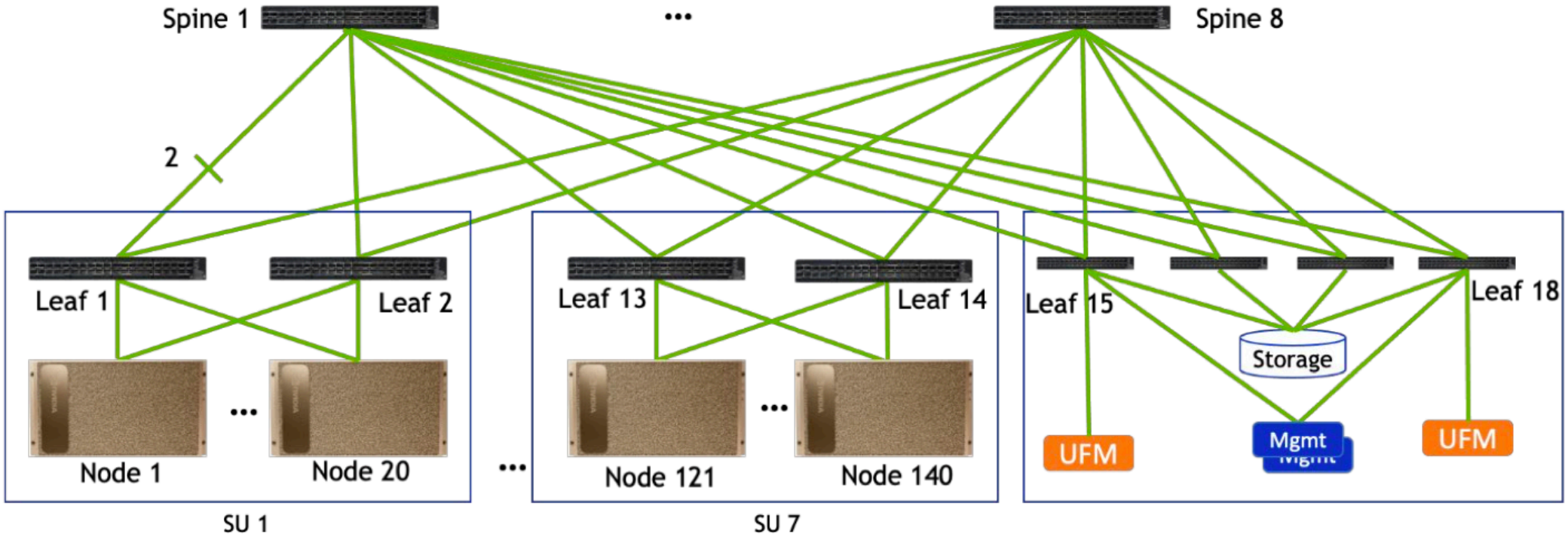
Figure 4. Compute fabric topology for a 140-node DGX SuperPOD



大模型业务全流程：AI 集群建设

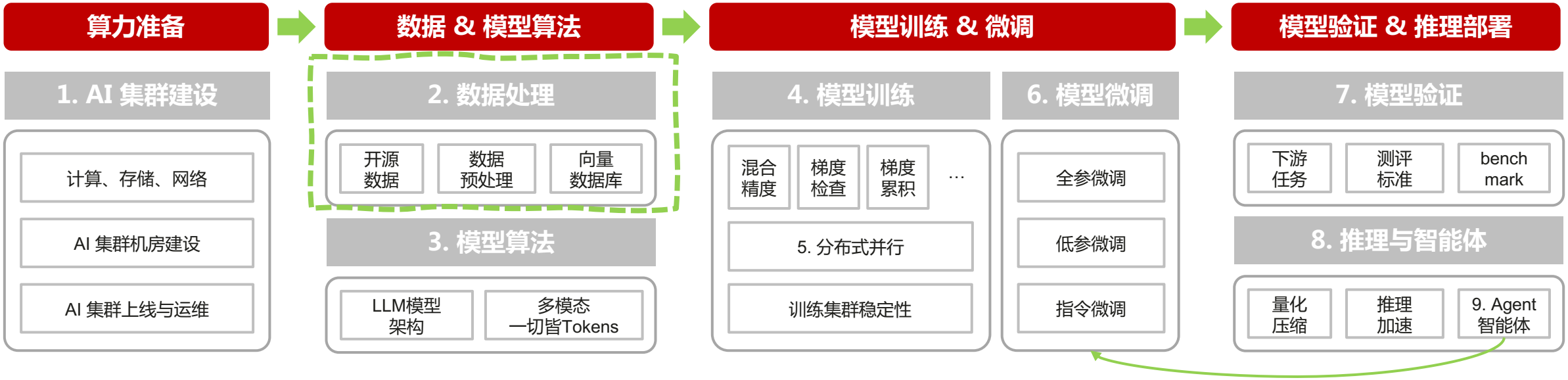
提供什么样的存储结构，才能适应大模型的发展和训练、推理方式？

Figure 7. Storage fabric topology for 140-node system



2. 数据处理

大模型业务全流程

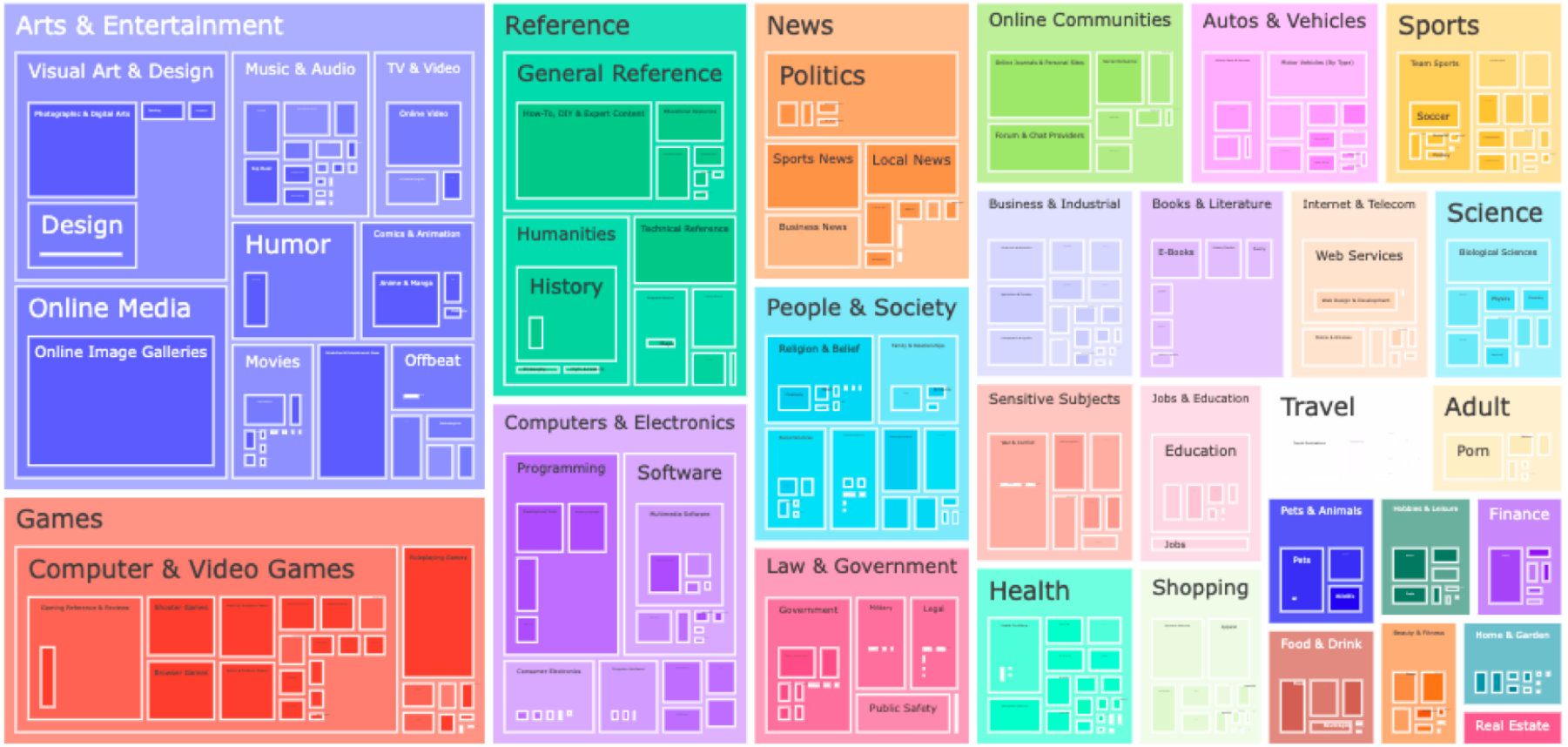


大模型不仅需要 LLM 算法，同时需要提供

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

2. 数据处理

- 大模型需要什么样的数据？以什么方式组织和组成数据？业界提供了哪些开源大数据集？



2. 数据处理

- LLM 没有数据预处理？Tokens 是通过什么算法生成、又如何从词到 Tokens？

Corpus	Prompt	PaLM 540B continuation
Wikipedia	the mother-tongue of 89.9% of the population, followed by Santali (8.1%), Kurmali Thar (1.1%), Hindi (0.5%) and Telugu (0.1%). According to the	West Bengal Official Language Act 1961 and the West Bengal Official Language (Amendment Act) 2012, the Bengali language is to be used for official purposes in the whole of West Bengal. In addition to Bengali, the Nepali

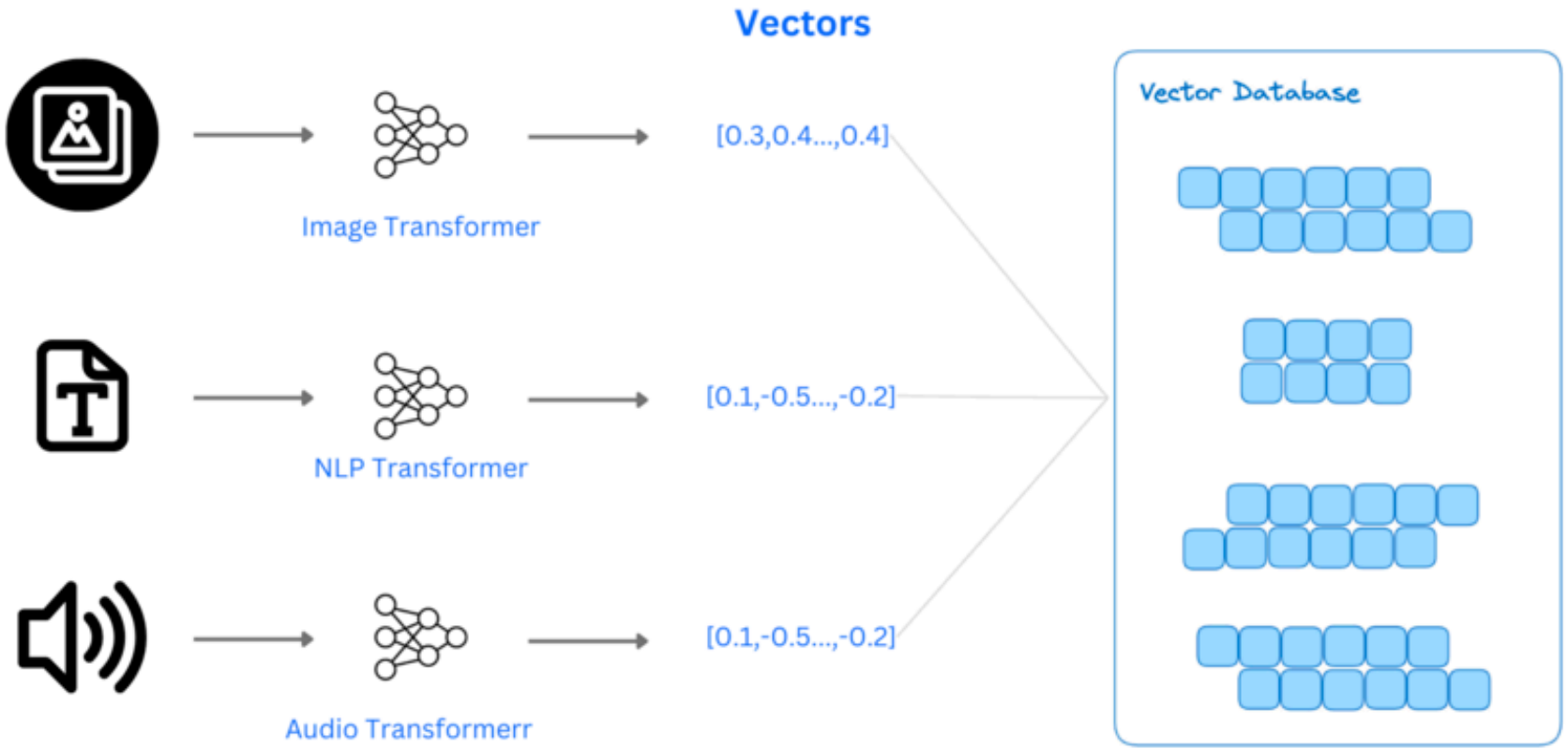
Code	[...] !address.isEmpty();
	}
	
	void EditAddressDialog::accept()
	{
	if(!model)
	return;
		if(!saveCurrentRow())
		{
		switch(model->getEditStatus())
		{
		case
		AddressTableModel::DUPLICATE_ADDRESS:
		QMessageBox::warning(this, windowTitle(),
		tr(“”The

News	Alexei Navalny arrived in Germany on a flight from Russia on Wednesday night, according to media reports. Yulia Navalnaya landed at Frankfurt airport after flying from Moscow on Wednesday evening, according to Der Spiegel, the German magazine. It is unclear what prompted Ms Navalnaya	to leave Russia, but associates stressed to the magazine that her departure was temporary. Her husband, a vocal critic of the Kremlin, was flown to Germany last summer after being poisoned in Siberia with what many Western countries said was a military-grade nerve agent.
Web	gonna get in your head somehow.” The man had said “Most human telepaths are adherents of Gaia” -The Interrogator made a mental note of the name, before allowing the thoughts to wash over him again- “But there are a fair number	of mutants who possess the ability. Not to mention the Zzyth, there seem to be a larger number of telepaths among them than among humanity, leading us to think they have either bred selectively for the trait, or their Pantheon is freer with
Code	/** * Copyright 2015 Google Inc. All Rights Reserved. * * Licensed under the Apache License, Version 2.0 (the ””License””); * you may not use this file except in compliance with the	License. * You may obtain a copy of the License at * * http://www.apache.org/licenses/LICENSE-2.0 * * Unless required by applicable law or agreed to in writing, software

ISO Code	Language	Tokens(B)	Percentage	ISO Code	Language	Tokens (B)	Percentage
en	English	578.064	77.984%	sd	Sindhi	0.149	0.020%
de	German	25.954	3.501%	haw	Hawaiian	0.140	0.019%
fr	French	24.094	3.250%	pa	Punjabi	0.138	0.019%
es	Spanish	15.654	2.112%	tg	Tajik	0.138	0.019%
und	Unknown	12.064	1.628%	tt	Tatar	0.137	0.019%
pl	Polish	10.764	1.452%	mk	Macedonian	0.135	0.018%
it	Italian	9.699	1.308%	kk	Kazakh	0.134	0.018%
nl	Dutch	7.690	1.037%	ru-Latn	Russian (Latin)	0.134	0.018%
sv	Swedish	5.218	0.704%	hmn	Hmong	0.132	0.018%
tr	Turkish	4.855	0.655%	te-Latn	Telugu (Latin)	0.123	0.017%
pt	Portuguese	4.701	0.634%	lv	Latvian	0.119	0.016%
ru	Russian	3.932	0.530%	fy	Frisian	0.111	0.015%
fi	Finnish	3.101	0.418%	is	Icelandic	0.110	0.015%
cs	Czech	2.991	0.404%	sq	Albanian	0.108	0.015%
zh	Chinese	2.977	0.402%	af	Afrikaans	0.108	0.015%
ja	Japanese	2.832	0.382%	ml	Malayalam	0.108	0.015%
no	Norwegian	2.695	0.364%	ja-Latn	Japanese (Latin)	0.107	0.014%
ko	Korean	1.444	0.195%	kn	Kannada	0.106	0.014%
da	Danish	1.387	0.187%	bg-Latn	Bulgarian (Latin)	0.089	0.012%
id	Indonesian	1.175	0.159%	mt	Maltese	0.086	0.012%
ar	Arabic	1.091	0.147%	mg	Malagasy	0.081	0.011%
ceb	Cebuano	1.012	0.137%	uz	Uzbek	0.080	0.011%
uk	Ukrainian	0.890	0.120%	ig	Igbo	0.079	0.011%
vi	Vietnamese	0.726	0.098%	so	Somali	0.076	0.010%
ps	Pashto	0.638	0.086%	am	Amharic	0.076	0.010%
ca	Catalan	0.636	0.086%	ga	Irish	0.076	0.010%
hu	Hungarian	0.555	0.075%	ne	Nepali	0.075	0.010%
ro	Romanian	0.510	0.069%	mr	Marathi	0.075	0.010%
la	Latin	0.482	0.065%	my	Burmese	0.069	0.009%
cy	Welsh	0.481	0.065%	si	Sinhalese	0.069	0.009%
iw	Hebrew	0.477	0.064%	ku	Kurdish	0.069	0.009%
fa	Persian	0.437	0.059%	mr-Latn	Marathi (Latin)	0.067	0.009%
bs	Bosnian	0.427	0.058%	ml-Latn	Malayalam (Latin)	0.066	0.009%
co	Corsican	0.392	0.053%	gu-Latn	Gujarati (Latin)	0.064	0.009%
ht	Haitian Creole	0.376	0.051%	ky	Kyrgyz	0.064	0.009%
sr	Serbian	0.373	0.050%	el-Latn	Greek (Latin)	0.063	0.008%
el	Greek	0.366	0.049%	mn	Mongolian	0.061	0.008%
eo	Esperanto	0.348	0.047%	km	Khmer	0.060	0.008%
be	Belarusian	0.329	0.044%	ckb	Unknown	0.055	0.007%

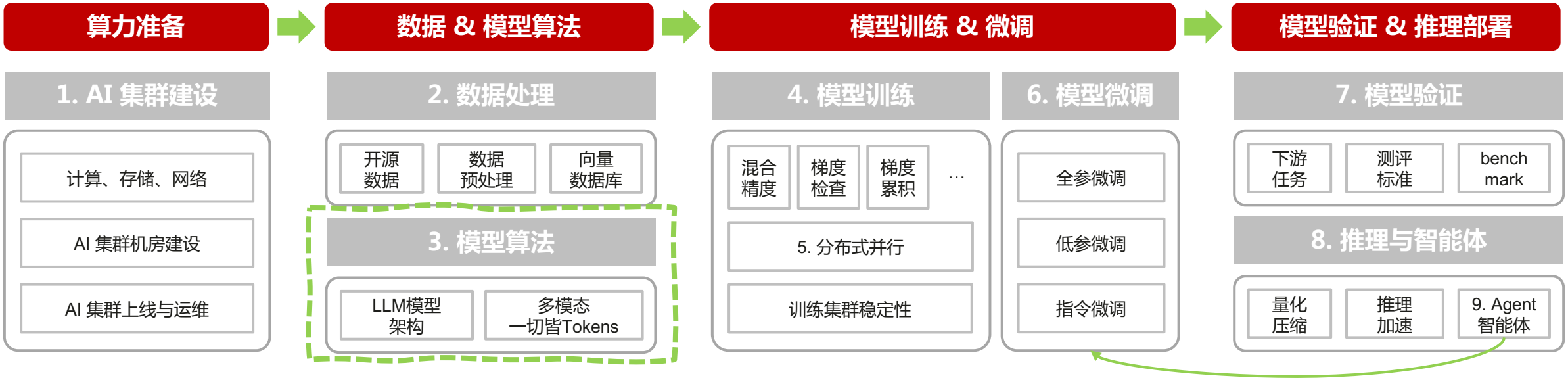
2. 向量数据库

- 向量数据库作为 AI 时代的数据基座，百 & 万亿级向量检索的向量数据库如何构建？



3. 大模型算法

大模型业务全流程

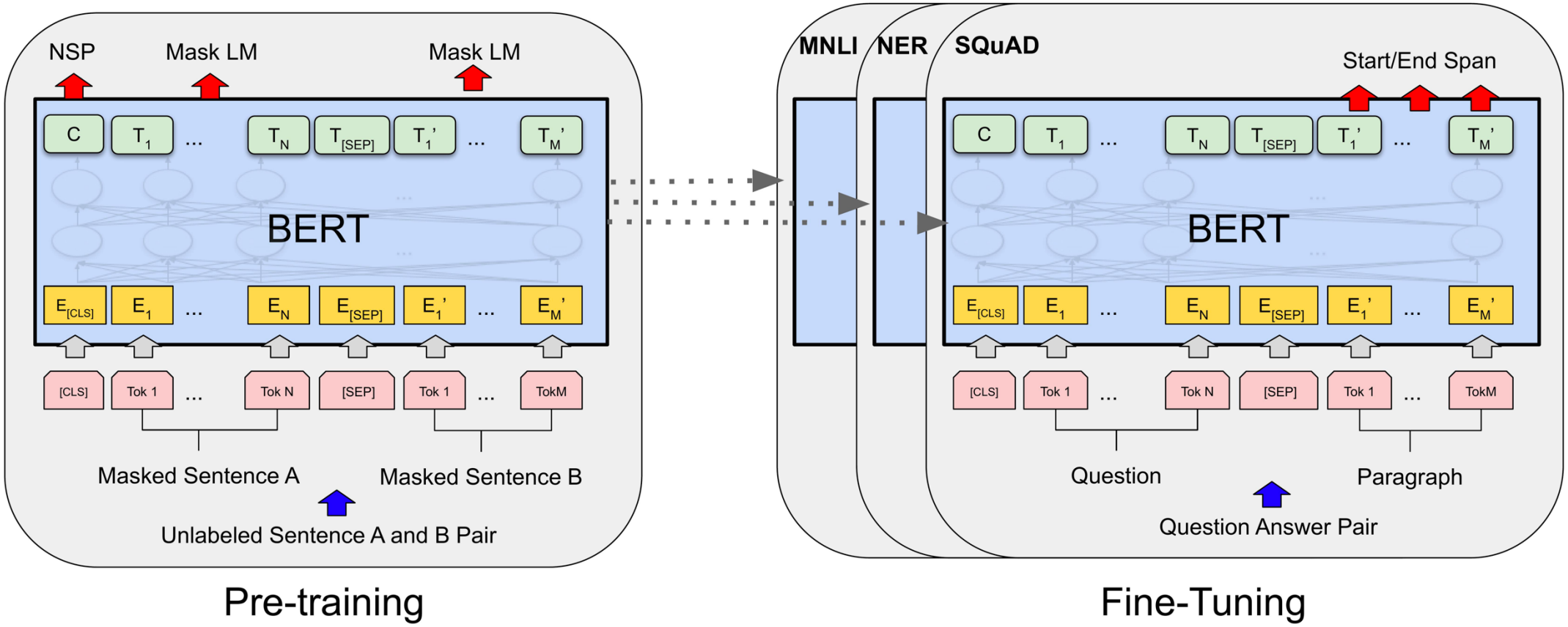


大模型不仅需要 LLM 算法，同时需要提供

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

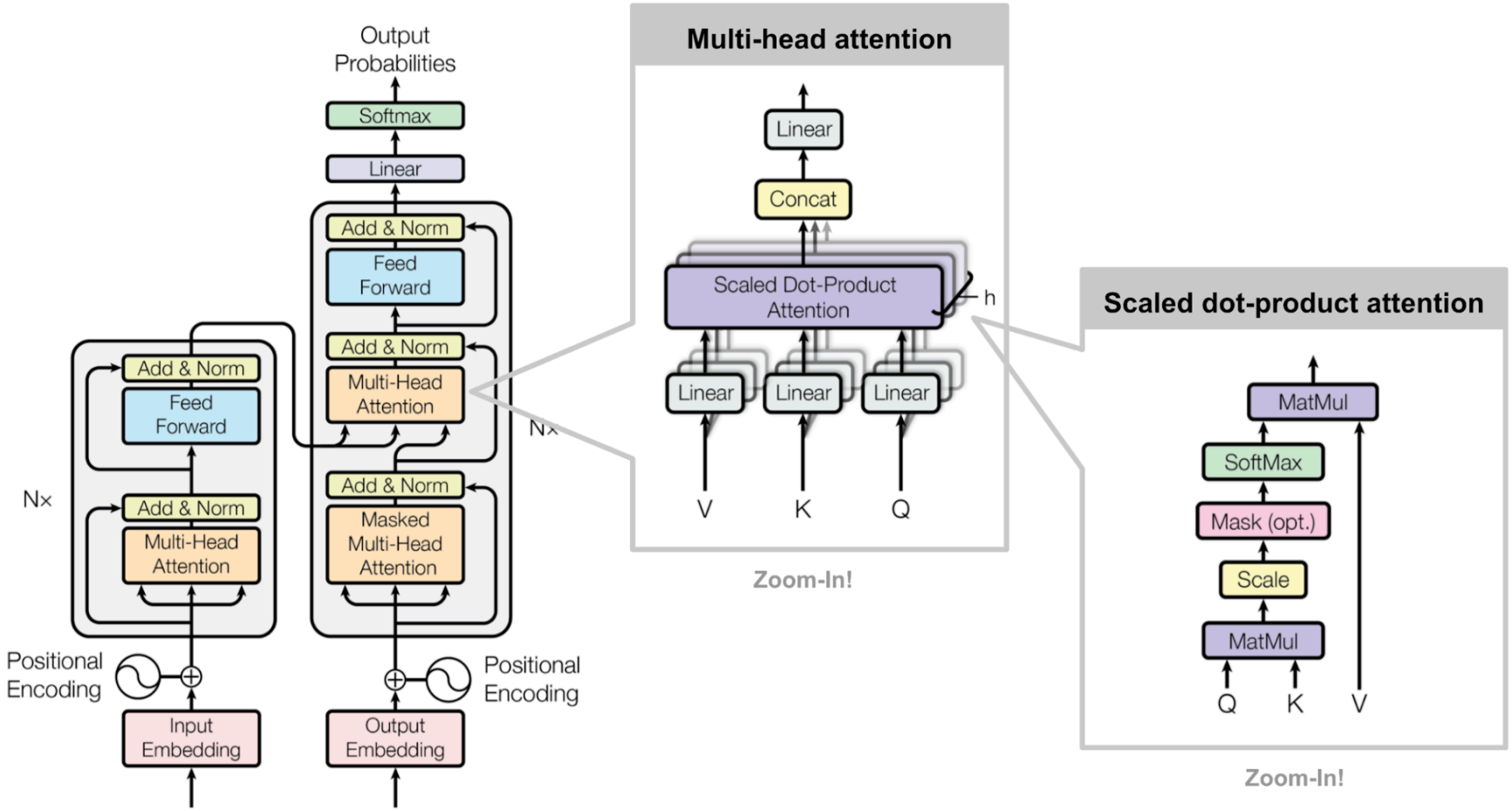
3. 大模型算法

- 传统 NLP 模型和小模型有什么缺陷，引入了预训练模型，从而演进到 LLM 大模型？



3. 大模型算法

- LLM 模型架构和原理，为什么说 Attention is all you need？大模型又是通过 Transform 如何组成？

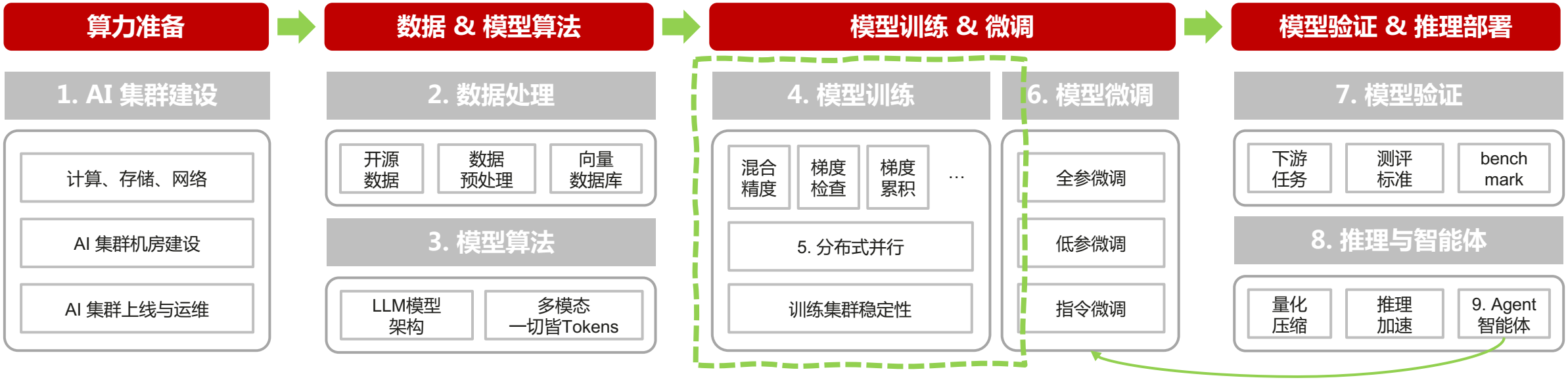


3. 大模型算法

- Transformer 如何统一一切？多模态大模型在不同模态之间， 如何组织不同类型任务和数据？

4. 大模型训练

大模型业务全流程

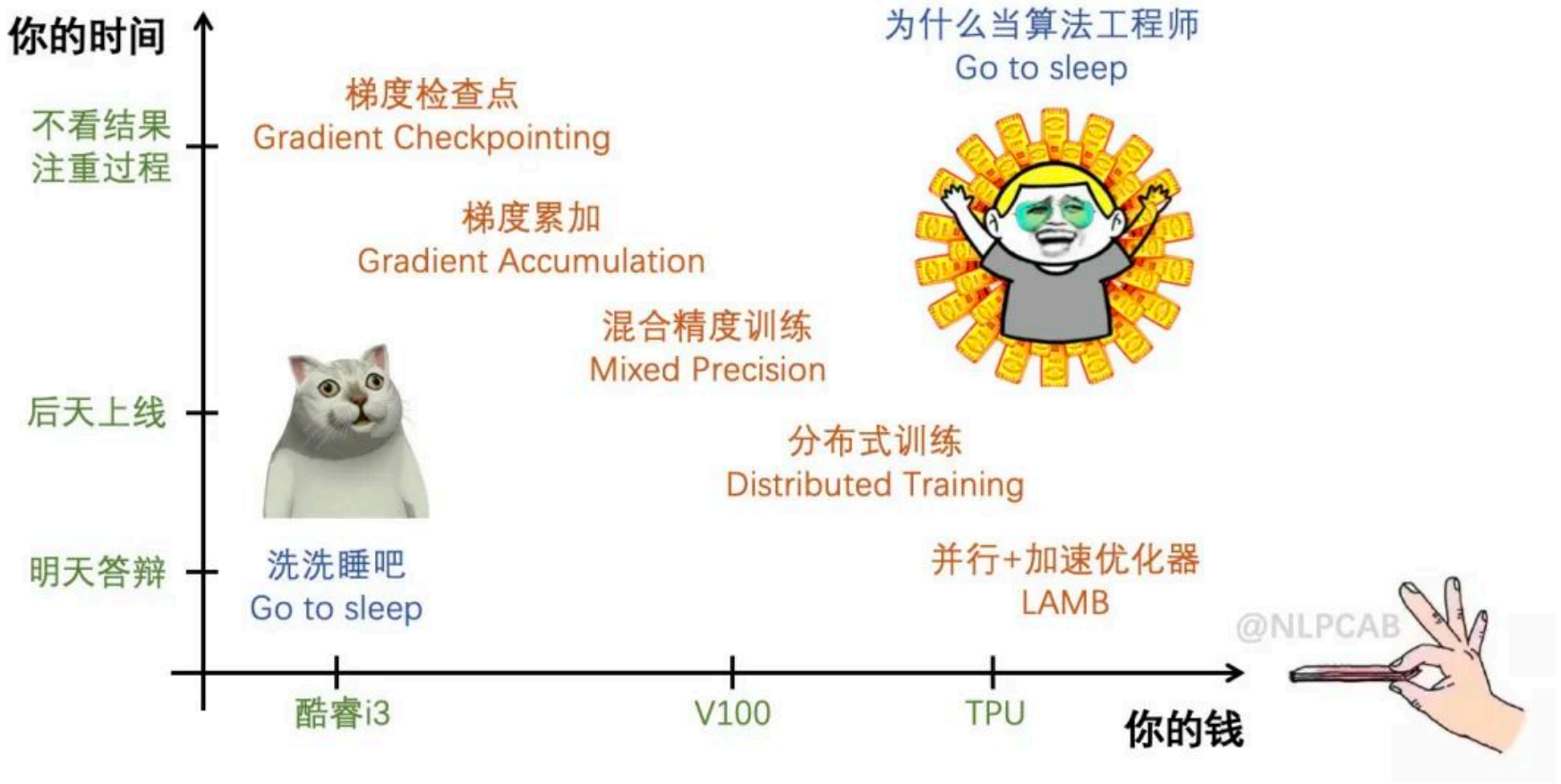


大模型不仅需要 LLM 算法，同时需要提供

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

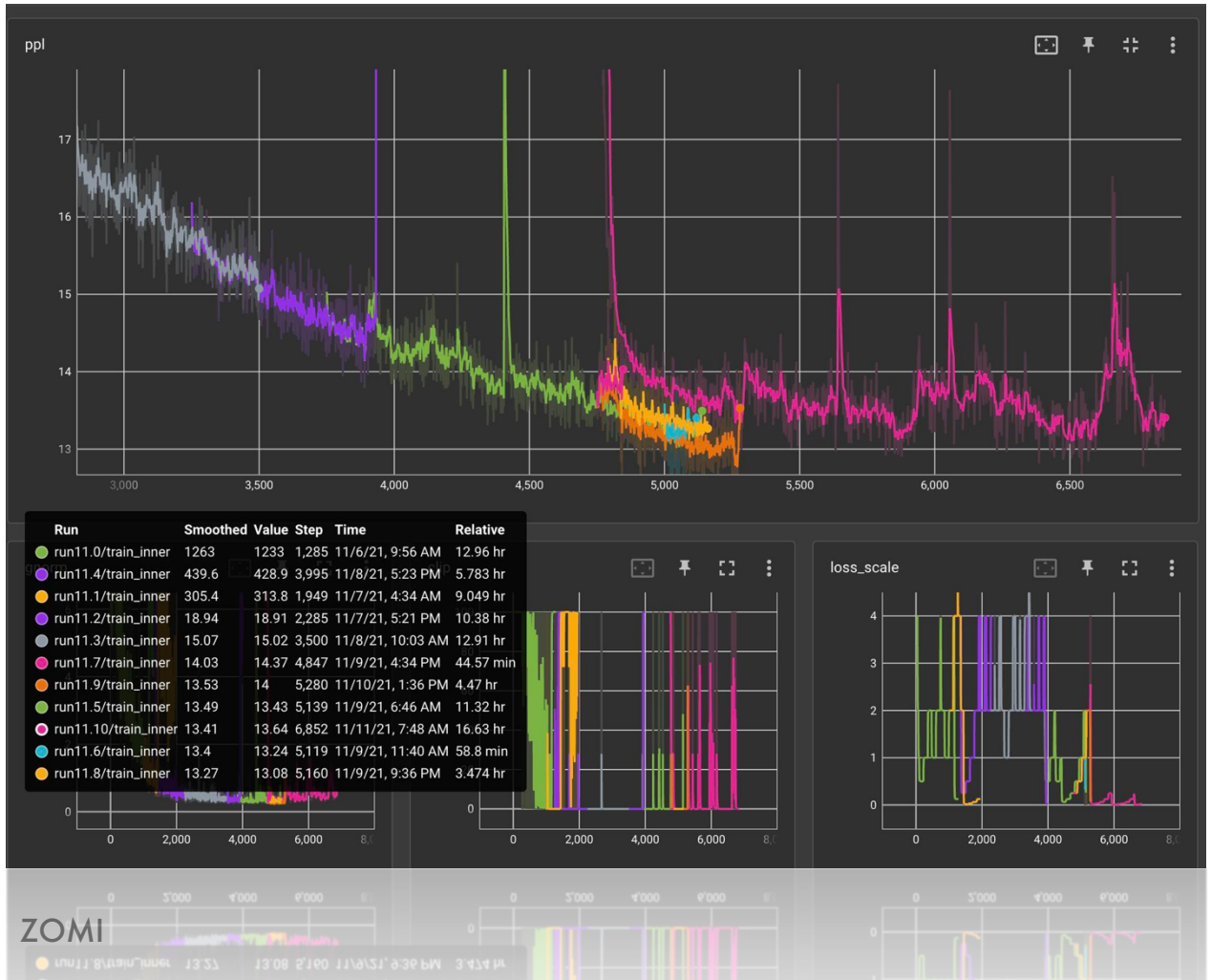
4. 大模型训练

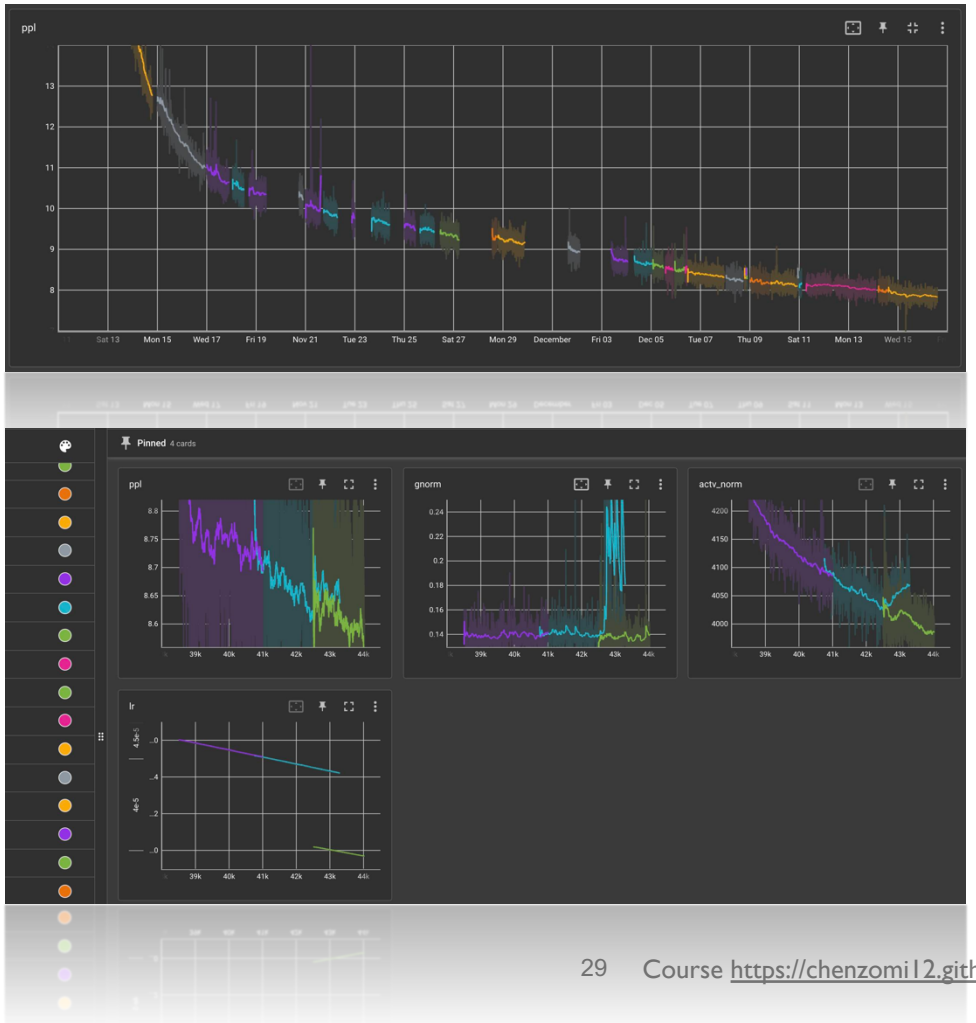
- 大模型训练在时间和算力之间怎么平衡？个人怎么玩转大模型？大模型时代“我”还有机会吗？



4. 大模型训练

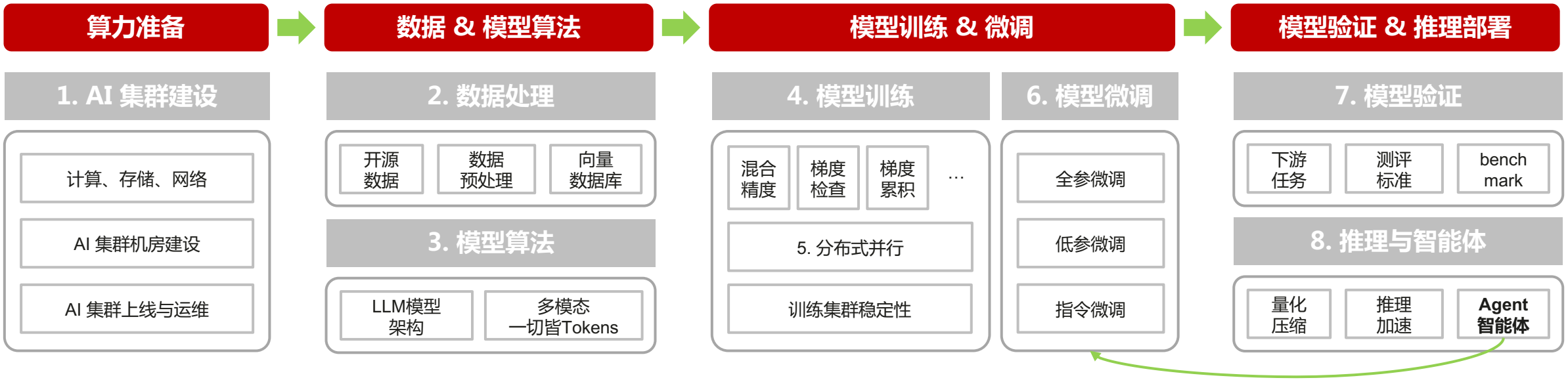
- 为什么单卡跑得好好的，集群几个小时断一次，几天崩一次，如此痛苦？如何搞定长稳？





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.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem