DeepSeek模型优势：算力、成本角度解读

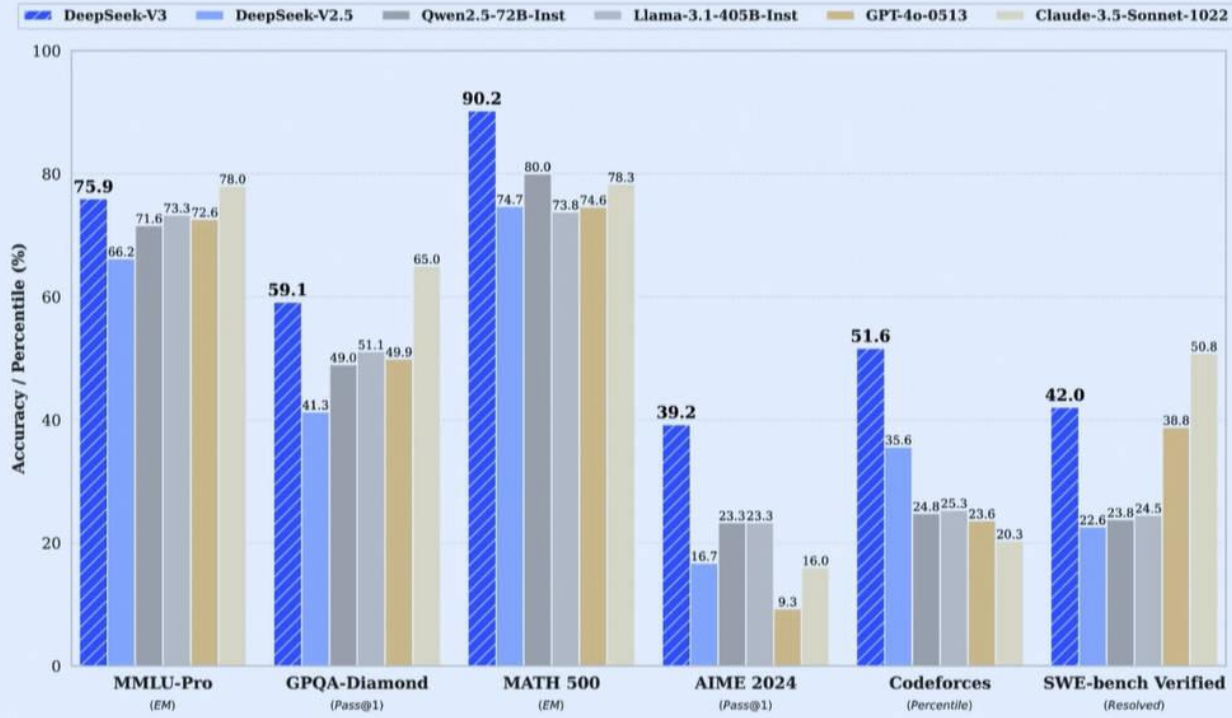
浙江大学计算机学院

浙江大学人工智能协同创新中心

王则可

2025年2月

DeepSeek优势



- DeepSeek的优势：系统感知的算法创新（量化基因）
  - 算法：霸榜，有创新（MLA、特定MoE）
  - 系统：低成本、高性能

算力的基本概念

- 什么算力?“对信息数据进行计算，实现目标结果的能力”
  - 传统算力：信息计算力
  - 现代算力：信息计算力、数据存储力、网络运载力

算力的发展

大脑

草绳、石子 ►

算盘、算筹 ►

计算器、计算机

- 原生算力：大脑（可处理复杂逻辑，但不能高速处理简单运算）
- 外部算力工具：
  - 草绳、石子
  - 算盘
  - 计算机：算力提供者（可高速简单运算，不能处理复杂逻辑）

计算机算力的发展

大型机时代
1940-1980

PC时代
1980-2000

云计算时代
2000-2020

人工智能时代
2020-

- 大型机时代：数字化未开始，算力需求潜力未发掘
- PC时代：一个应用只需一台电脑，算力够
- 云计算时代：应用需要超过一台机器的算力，算力基本够
- 人工智能时代：算力开始不足，需大量高性能AI加速器

人工智能大模型算力估计

- 人工智能大模型算力估计
  - 1，数据量（D）＞15＊模型参数量（N）
    - 万亿模型(N) = $1000*10^9 = 10^{12}$
    - 数据量(D) > $15 * 10^{12} = 1.5* 10^{13}$
  - 2，计算次数 C ≈ 6 * N * D
    - 万亿模型计算次数 C ≈ 6 * N * D ≈ $1.5 * 10^{25}$

OpenAI. “Scaling Laws for Neural Language Models”, 2020

人工智能计算平台成本估计

算力　　存力　　运力

	算力 (每秒)	显存	运力	生态	政策风险	成本
华为910B	320T = $3.2*10^{14}$	32GB	240 GB/s	较好	无	**12万**
英伟达H800	**1000T = 10^{15}**	**80GB**	**900 GB/s**	好	有	25万

- 万亿大模型预训练系统成本估计
  - 条件：计算量 $C \approx 6 * N * D \approx 1.5 * 10^{25}$
  - 最低时间、成本估计

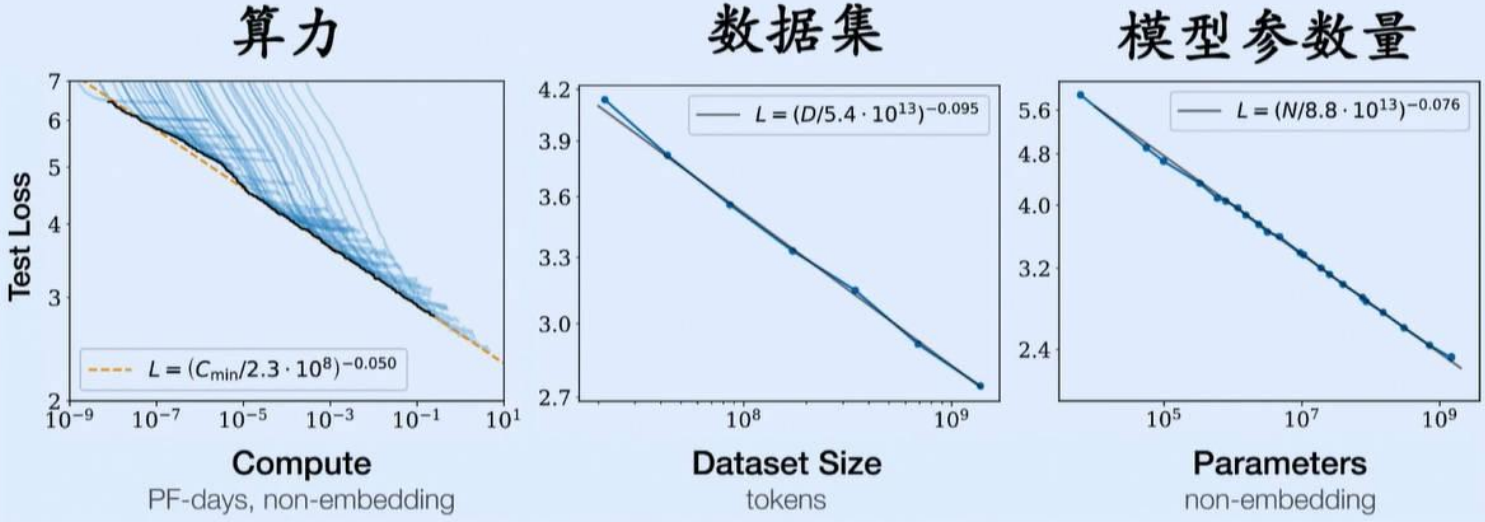
人工智能计算平台成本估计

算力　　存力　　运力

	算力 (每秒)	显存	运力	生态	政策风险	成本
华为910B	320T = $3.2*10^{14}$	32GB	240 GB/s	较好	无	**12万**
英伟达H800	**1000T = 10^{15}**	**80GB**	**900 GB/s**	好	有	25万

- 万亿大模型预训练系统成本估计
  - 条件：计算量 $C \approx 6 * N * D \approx 1.5 * 10^{25}$
  - 最低时间、成本估计
    - 单H800（25万）：**$1.5* 10^{10}$ 秒** （174000天）
    - 1000张H800（2.5亿）：**$1.5* 10^{7}$ 秒** （174天）

大模型指导法则Scaling Law：富则火力覆盖



- 大模型扩展规律（资本非常喜欢确定性故事）
  - 算力：算力越大（x轴），模型效果越好（Test Loss小）
  - 数据集：数据集越大（x轴），模型效果越好
  - 模型参数：参数越多（x轴），模型效果越好

国际人工智能企业OpenAI的商业模式

- OpenAI商业模式（循环以下四步）
  - 1，华尔街融资
    - 例子：2019-21年融资20亿美元
  - 2，购买最新GPU
    - 例子：购买2.5万A100 GPU （英伟达挣钱）
  - 3，用最新GPU训练性能领先的大模型
    - 例子：2023年出ChatGPT，垄断市场（290亿美元估值）
  - 4，用训练的GPU给客户提供高质量模型服务
    - 例子：营收小、整体亏钱

2025年特朗普的“星际之门”为OpenAI筹5000亿美元AI基础设施！

美国限制中国AI发展的策略

- 国内人工智能商业模式（循环以下四步）
  - 1，国内融资（亿美金）
    - 可行性分析：资金没问题，尤其优质生产力领域
  - 2，购买最新GPU
    - 可行性分析：美国可以发禁令
  - 3，用GPU训练性能领先的大模型
    - 可行性分析：国内AI人才没问题
  - 4，用训练的GPU给客户提供高质量模型服务
    - 可行性分析：国内做工业化低成本有绝对优势

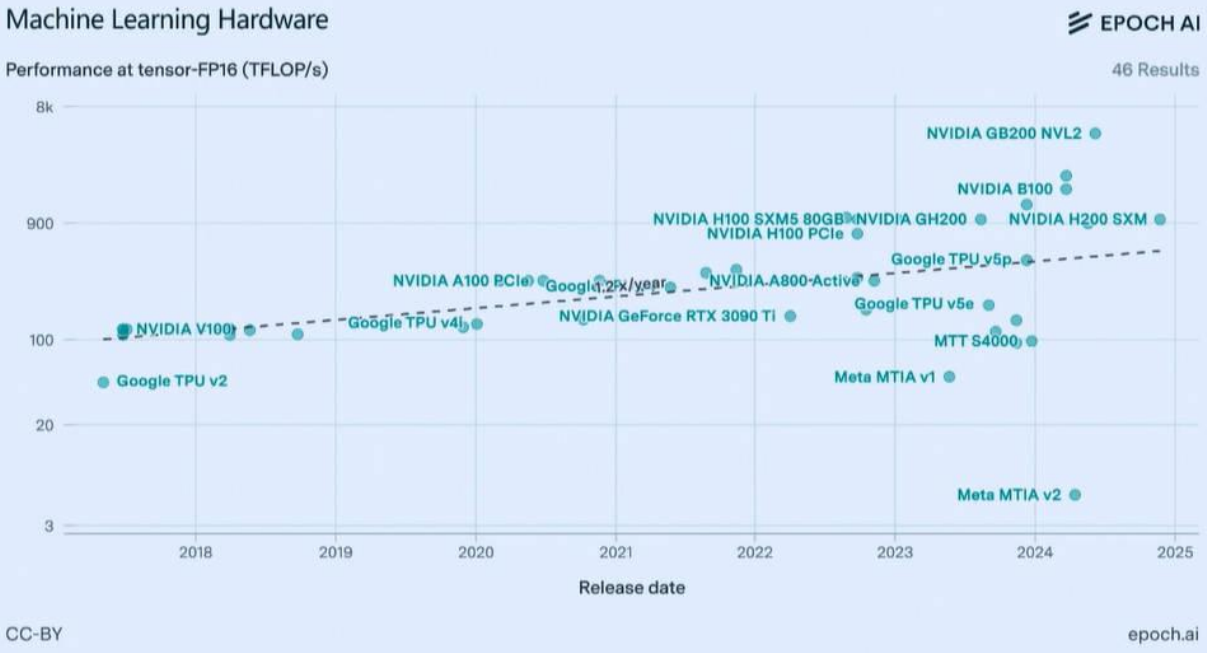
时代背景：算力卡脖子

- 美国政府对我国的禁令
  - 现成成熟算力：2023年禁止出口高端AI芯片
    - A100、H00、H800、A800 等数据中心GPU
  - 运力：2022年限制AI加速器的互联带宽
  - 算力：2024年禁止台积电代工7nm工艺的国内芯片
  - 存力：2024年禁止HBM芯片
  - 光刻机：2024年限制荷兰ASML出口7nm光刻机到中国

卡脖子后果：国内AI优质算力有差距

浙江大学 ZHEJIANG UNIVERSITY

	算力 (每秒)	显存	运力	生态	政策风险	成本
华为910B	320T = $3.2*10^{14}$	32GB	240 GB/s	较好	无	12万
英伟达H800	1000T = 10^{15}	80GB	900 GB/s	好	有	25万



DeepSeek等国内大模型的“上甘岭”时刻

浙江大学 ZHEJIANG UNIVERSITY

	“大模型”	“上甘岭”
国际	OpenAI 大资金、大算力、大模型	 范弗利特弹药量(地毯轰炸)
国内	deepseek Qwen AI算法与系统协同深度优化	 反斜面坑道(战术穿插)

DeepSeek V3 公开的单次极低预训练成本

	发布时间	GPU时（小时）	训练成本（美元）
Llama 3.1	2024年7月	$3.1*10^7$	$6.2*10^7$
DeepSeek v3	2024年12月	**$2.8*10^6$**	**$5.6*10^6$**

- DeepSeek全部训练单次成本：5,576,000 美元
  - 单张H800 GPU 每小时租赁成本：2 美元

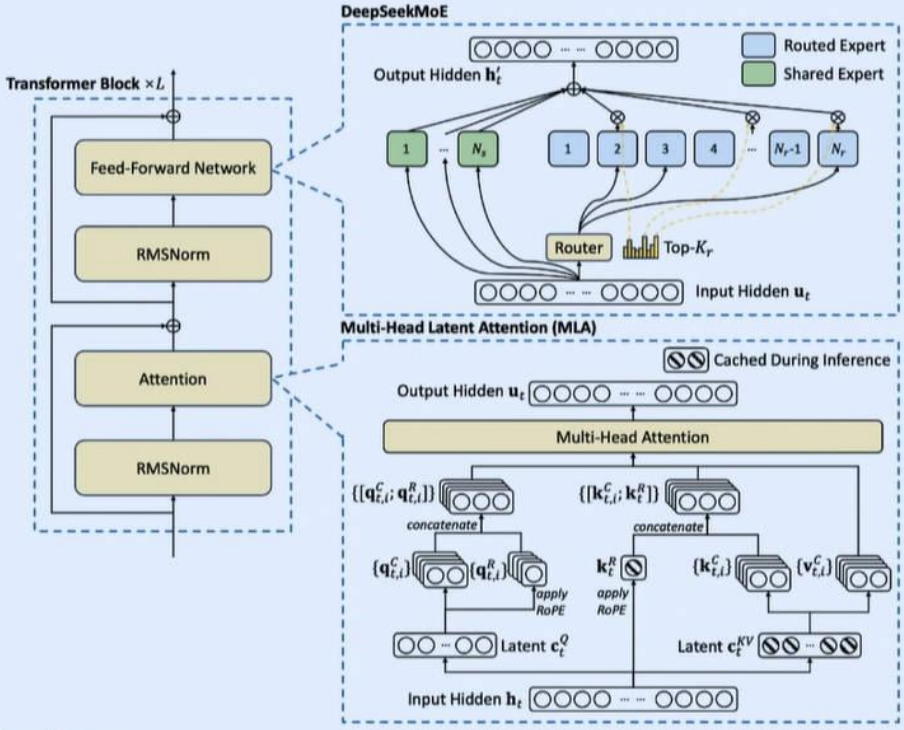
DeepSeek发展历程：穷则战术穿插

浙江大学 ZHEJIANG UNIVERSITY

模型/指标	DeepSeek V1	DeepSeek V2	DeepSeek V3	Llama 3.1
发布时间	2024年1月	2024年6月	2024年12月	2024年7月
训练 Token	2 T	8.1 T	14.8 T	15T
模型规模	7B、67B	236B / 激活21B	671B / 激活37B	405B
MoE模型	稠密	MoE 2+160	MoE 1+256	稠密
注意力技术	GQA	MLA	MLA	N.A
上下文长度	4K	128K	128K	128K
训练成本 (GPU Hours)	300.6K	172.8K	**2.788 M**	**30.84 M**

DeepSeek v3 模型参数

L = 61 层

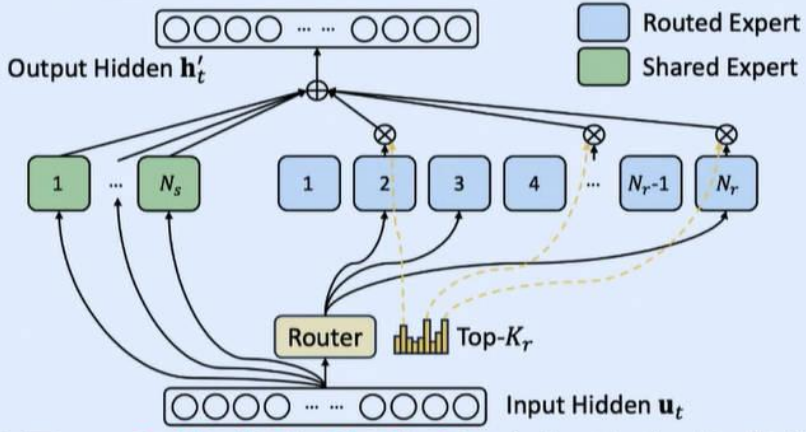


MoE: 1 共享专家+ 256 路由专家

MLA: 低秩压缩

- DeepSeek V3 模型参数?
  - 671B 参数 （GPT-3：175B、GPT-4：1.76T?）
  - 每个 token 激活 37B参数(~5.5%)，降低计算量

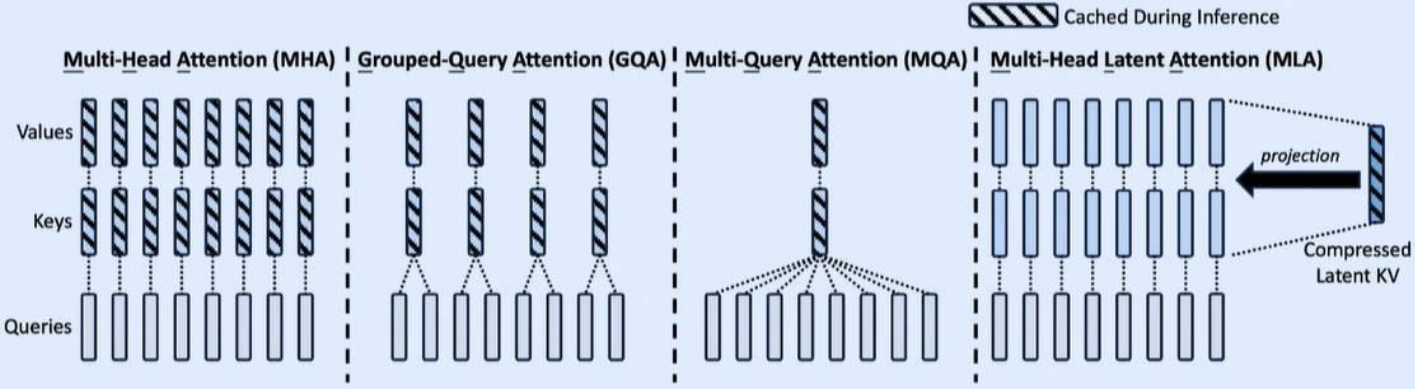
核心技术DeepSeekMoE：显著减少计算量



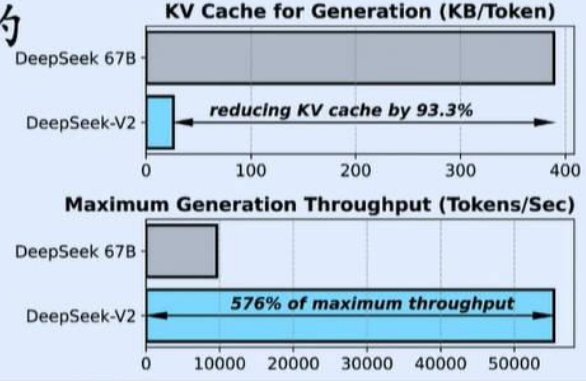
- 核心技术DeepSeekMoE：显著减少计算量（穷则战术穿插）
  - 针对美国的算力禁令
  - 核心思想：1 共享专家 +256 路由专家，激活 8 个路由专家
    - 共享专家：捕获通用知识、降低知识冗余
    - 路由专家：量大、细粒度、灵活组合、方便知识表达
  - 结果：每个Token只要过360亿参数（Llama 3.1要4050亿参数）

DeepSeek. "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models", 2024

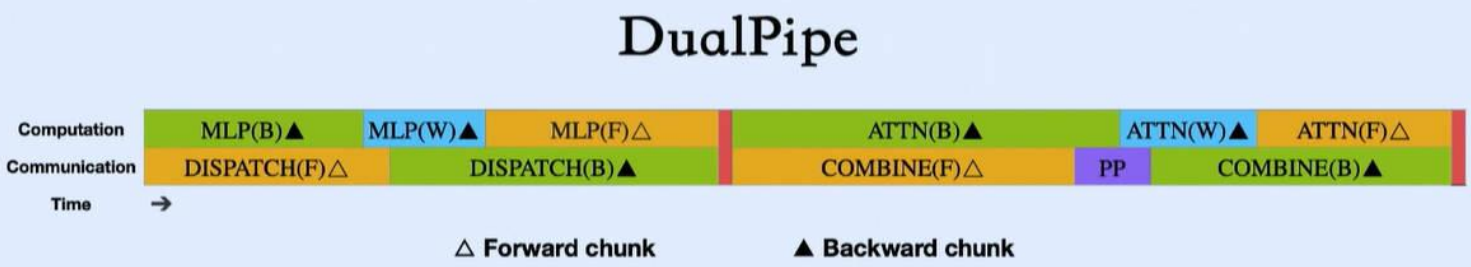
核心技术MLA：Multi-Head Latent Attention



- MLA：少许计算量换HBM空间（穷则战术穿插，已开源）
  - 针对美国的HBM芯片禁令(AI算力严重依赖高性能内存)
  - 核心思想：低秩压缩 KV，显著降低推理时 KV cache 的存储空间需求
  - 结果：KV Cache 使用降低 93.3%
    - 推理性能：显著提升
    - 推理成本：显著降低



系统核心技术：自研轻量级框架



- DeepSeek其它方面的性能方面优化
  - 自研轻量级框架（允许系统极致性能优化）
  - FP8训练（提升算力密度）
  - DualPipe（通信、计算重叠度高）
  - PTX优化绕开CUDA护城河 （单独解读）

DeepSeek有无绕开CUDA护城河?

浙江大学 ZHEJIANG UNIVERSITY

DeepSeek论文

selects only 8 routed experts in practice, it can scale up this number to a maximum of 13 experts (4 nodes × 3.2 experts/node) while preserving the same communication cost. Overall, under such a communication strategy, only 20 SMs are sufficient to fully utilize the bandwidths of IB and NVLink.

In detail, we employ the warp specialization technique (Bauer et al., 2014) and partition 20 SMs into 10 communication channels. During the dispatching process, (1) IB sending, (2) IB-to-NVLink forwarding, and (3) NVLink receiving are handled by respective warps. The number of warps allocated to each communication task is dynamically adjusted according to the actual workload across all SMs. Similarly, during the combining process, (1) NVLink sending, (2) NVLink-to-IB forwarding and accumulation, and (3) IB receiving and accumulation are also handled by dynamically adjusted warps. In addition, both dispatching and combining kernels overlap with the computation stream, so we also consider their impact on other SM computation kernels. Specifically, we employ customized PTX (Parallel Thread Execution) instructions and auto-tune the communication chunk size, which significantly reduces the use of the L2 cache and the interference to other SMs.

- PTX （Parallel Thread Execution）类英伟达汇编
  - 作用：C++抽象较高，无法表达GPU内部硬件特性，PTX指令控制 1）内存读写到L2、内存和 2）GPU内部硬件引擎
  - 个人猜测：GPU的内存一致性模型做的差，故GPU计算和通信的内存一致性只能用PTX指令来保证
  - 结论：没绕开，更依赖CUDA；但对国产硬件设计有作用

DeepSeek V3 未来展望

- DeepSeek 为代表的国内大模型咬住国外最先进大模型
  - 模型性能：不要指望全面优势，“城头变幻大王旗”
  - 成本：低（战术穿插）
  - 算力受限，近几年咬住会更难（大家宽容些）
- 突破工艺卡脖子，实现“战术穿插”+“火力覆盖”
  - 中芯国际等硬核大厂突破工艺卡脖子
  - 华为等算力公司提供高算力密度
- 个人预测AI竞赛结果
  - 以中国的工业化水平，站着把AI的钱给挣了。
  - “健身可以让SB跟你好好说话”→

    “突破模型、算力卡脖子可以让A国跟咋们好好说话”

本次讲座是科普性质，
有不严谨的地方敬请谅解！