AI大模型开启新一轮大国竞争， 半导体战略地位凸显

核心观点

- **AI正处史上最长繁荣大周期：** 在进入21世纪以来，在大数据和大算力的支持下，归纳统计方法逐渐占据了人工智能领域的主导地位，深度学习的浪潮席卷人工智能，人工智能迎来史上最长的第三次繁荣期， 至今仍未有结束的趋势。
- **OpenAI的“暴力美学”：大算力和大数据：** OpenAI 认为，通过独立延长模型训练时间、增加训练数据量或者扩大模型参数规模，预训练模型在测试集上的 Test Loss 都会单调降低，从而使模型效果越来越好。我们认为， 在 Scaling Law 的框架下，只要追加数据与算力，大模型的能力就能持续增强。 **对于OpenAI而言，目前大模型的最大限制是数据和算力的总量。**
- **大模型开启新一轮大国竞争，半导体成顶层博弈焦点：** 预训练大模型是现阶段人工智能的集大成者， 代表了统计学习流派的最高成就。在新一代技术未出现前，它将是人工智能研究和开发的最强武器。围绕大模型的研发和落地，中美之间已经展开了新一轮的竞争，美国已对华限制销售最先进的英伟达A100和H100 GPU 训练芯片。 **半导体作为AI算力核心， 将受到顶层高度关注，成为大国博弈的焦点之一。**
- **AI模型运算规模增长， 算力缺口巨大：** 基于大量数据训练、拥有巨量参数的AI预训练模型—GPT-3，引发了AIGC技术的质变， 从而诞生ChatGPT。然而， 预训练模型参数数量、训练数据规模将按照 300 倍/年的趋势增长， 现有算力距离AI应用存巨大鸿沟。运算规模的增长， 带动了对AI训练芯片单点算力提升的需求， 并对数据传输速度提出了更高的要求。
- **建议关注：**
- GPU ： 景嘉微、航锦科技，海光信息和未上市的地平线、黑芝麻、摩尔线程
- AI训练芯片：寒武纪、商汤(港股)、燧原科技(未上市)
- AI存力：兆易创新、北京君正、 东芯股份
- HBM：雅克科技、深科技
- 半导体大国重器：中芯国际、北方华创、中微公司
- **风险提示：** AI算法、模型存较高不确定性， AI技术发展不及预期； ChatGPT用户付费意愿弱，客户需求不及预期； 针对AI的监管政策收紧

目录

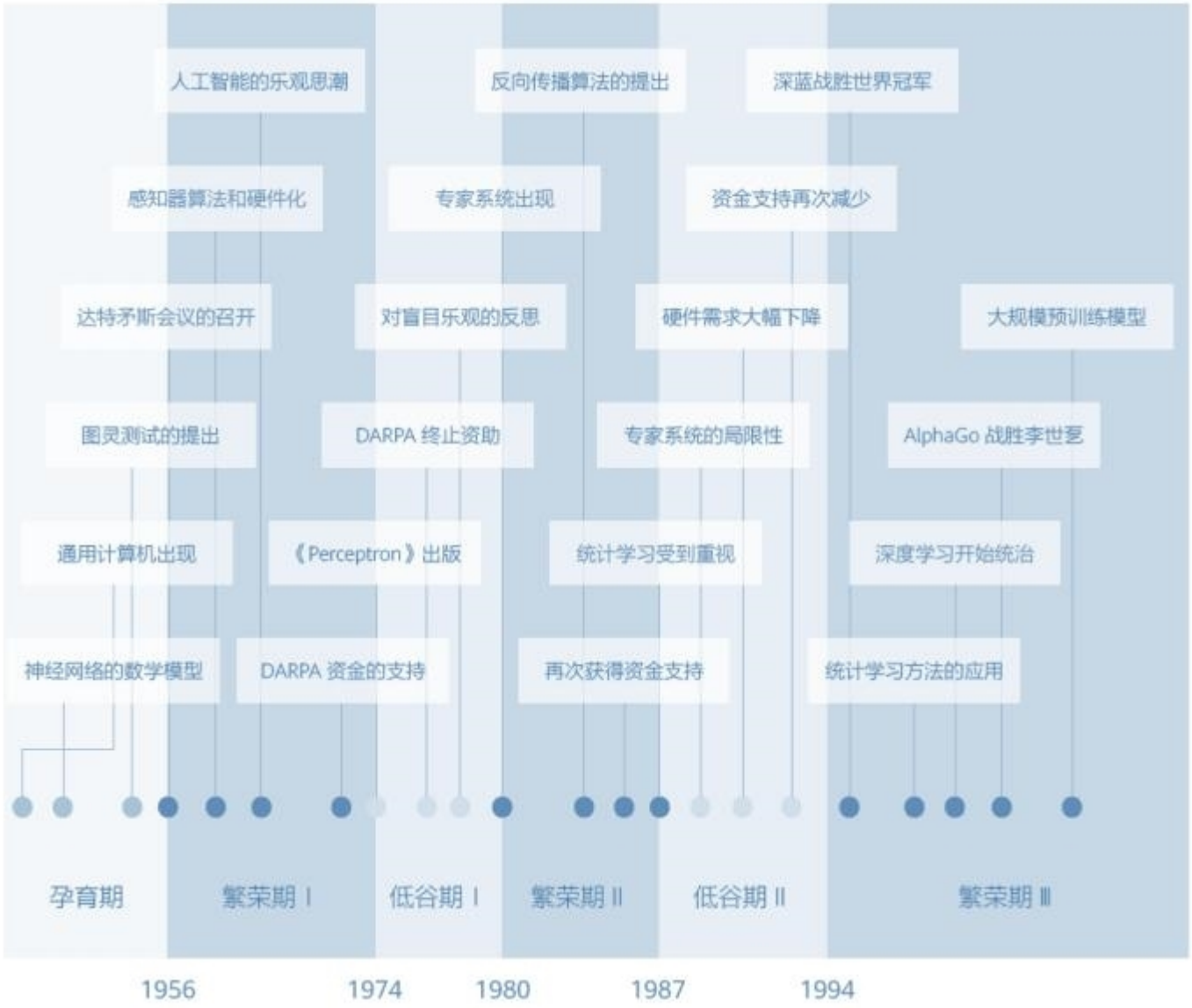
AI正处史上最长繁荣大周期

- 人工智能从1956 年被正式提出以来，经历了数十年的发展历程。人工智能诞生初期， 其研究主要分为三个流派，即逻辑演绎、归纳统计和类脑计算。

- 人工智能研究的三大流派各有优劣势。类脑计算流派的目标最为宏远，但在未得到生命科学的支撑之前，难以取得实际应用。归纳演绎流派的思考方式与人类相似，具有较强的可解释性。由于对数据和算力的依赖较少，归纳演绎流派成为人工智能前两次繁荣的主角。随着学界对人工智能困难程度的理解逐渐加深，数理逻辑方法的局限性被不断放大，并最终在第三次繁荣期中，逐渐让位于统计学习的“暴力美学”。

- **在进入21世纪以来，在大数据和大算力的支持下，归纳统计方法逐渐占据了人工智能领域的主导地位，深度学习的浪潮席卷人工智能， 人工智能迎来史上最长的第三次繁荣期，至今仍未有结束的趋势。**

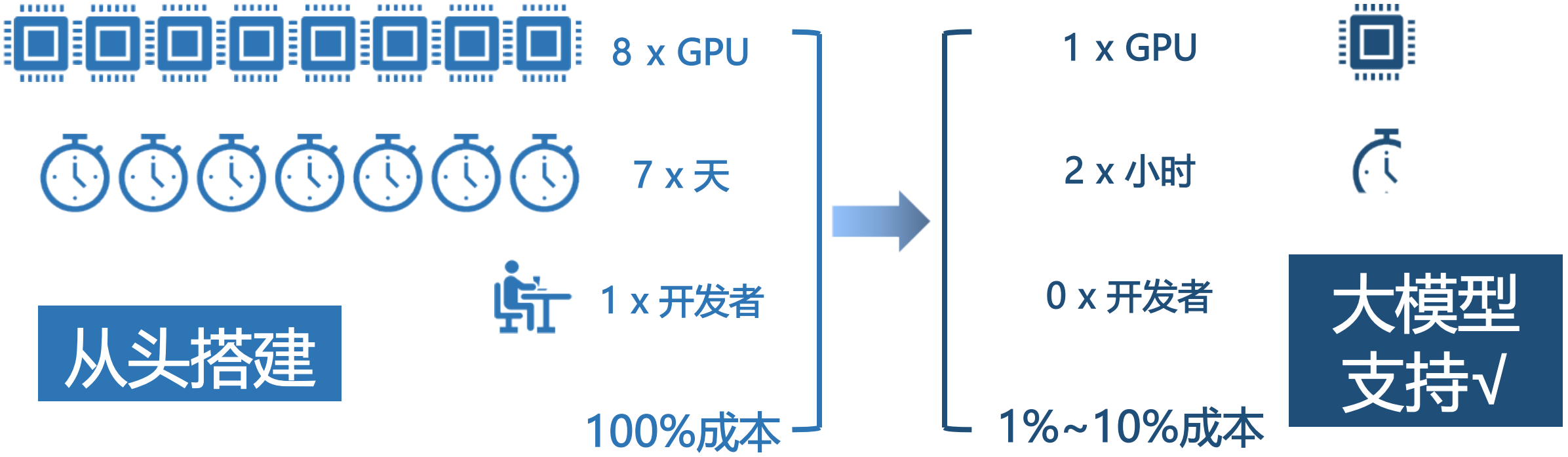
图：人工智能发展史



通用大模型加持，平民化AI普惠千行百业

- **深度学习依然受到统计学习的框架限制：特征抽取和模板匹配。** 相比于人类基于知识的推断， 这种方式无疑是低效的，因为对于任何新的概念乃至新的实体，算法都需要专门的训练数据来提供相关的信息。**在没有基础模型支撑的情况下， 开发者们必须从头开始完成收集数据、训练模型、调试模型、优化部署等一系列操作。** 对于人工智能开发者和垂直细分行业应用而言，都是重大的挑战。
- **预训练大模型降本增效， 将推动AI普惠千行百业。**预训练大模型加持下的人工智能算法(包括计算机视觉、自然语言处理等)，相比于普通开发者从头搭建的算法，精度明显上升、数据和计算成本明显下降，且开发难度大幅降低。

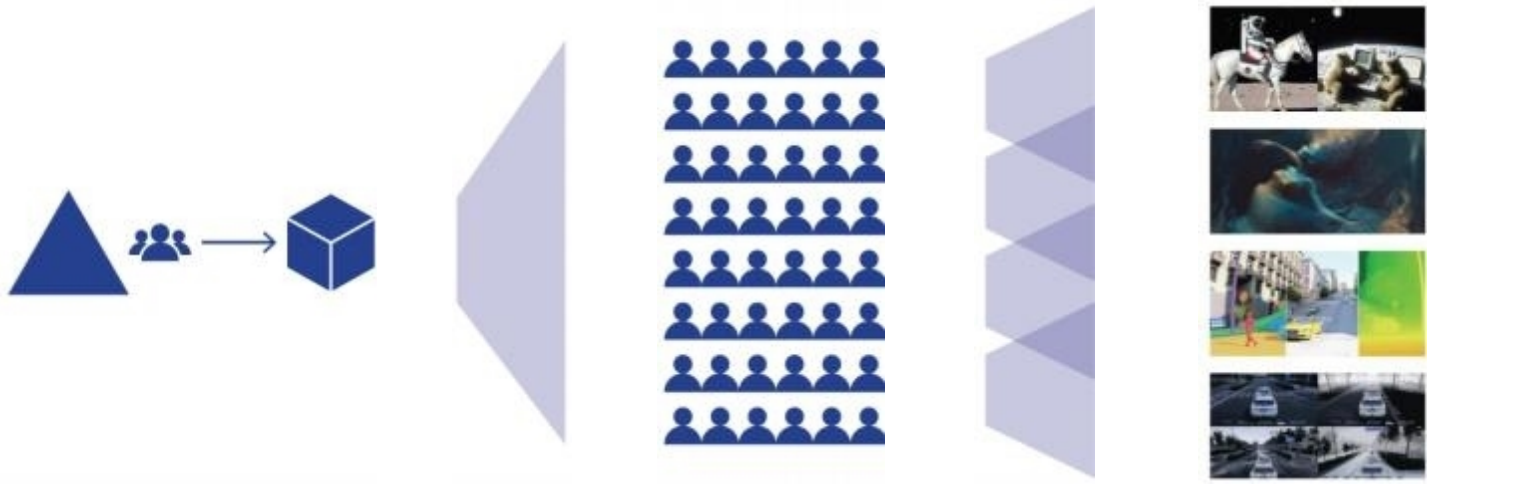
图：在100 张图像上训练基础物体检测算法，从头搭建 vs 大模型支持



GPT基础大模型驱动，引发AIGC范式革命

- **以ChatGPT为代表的AIGC应用在 2022 年的爆发，主要是得益于深度学习模型方面的技术创新。** 不断创新的生成算法、预训练模型、多模态等技术融合带来了 AIGC (AI Generated Content)技术变革，拥有通用性、基础性多模态、参数多、训练数据量大、生成内容高质稳定等特征的 AIGC 模型成为了自动化内容生产的“工厂”和“流水线”。
- **基础层是核心， GPT-3模型起关键支撑作用。** GPT-3一个大规模的通用语言模型， 已经在来自各种来源的大量文本数据上进行了训练。能够产生类似人类的反应，并可用于广泛的语言相关任务。
- ChatGPT基于目前较新的GPT-4模型版本进行研发，专注于自然语言对话，接受了更广泛的语言模式和风格培训，因此，能较GPT-4产生更多样化和微妙的响应。

图： AIGC产业架构



基础层	中间层	应用层
预训练模型	重直化、场景化、个性化模型	图像、语音和文字生成等各种各样的AIGC的应用
以预训练大规模模型为基础搭建的 AIGC 技术基础设施层。	在基础层上生成场景化、定制化、个性化的小模型，实现不同行业、垂直领域的流水线式部署	面向 C 端用户的文字、图片、音视频等内容生成服务

OpenAI的“暴力美学”：大算力和大数据

- **穷尽所有的测试数据和训练材料，AI就会呈现出恐怖的准确率。**OpenAI 意识到了“大”和“规模”的力量，沿着该路径狂飙，阅览了几乎所有互联网数据，并在超级复杂的模型之下进行深度学习。
- 2017-2019年，OpenAI 做出了有别于市场共识的关键决策，公司在Transformer 基础上押注大算力和大数据的“暴力美学”。并在GPT-3后迅速引入了人类反馈，让模型的语言前后逻辑更加明晰、有因果关联。

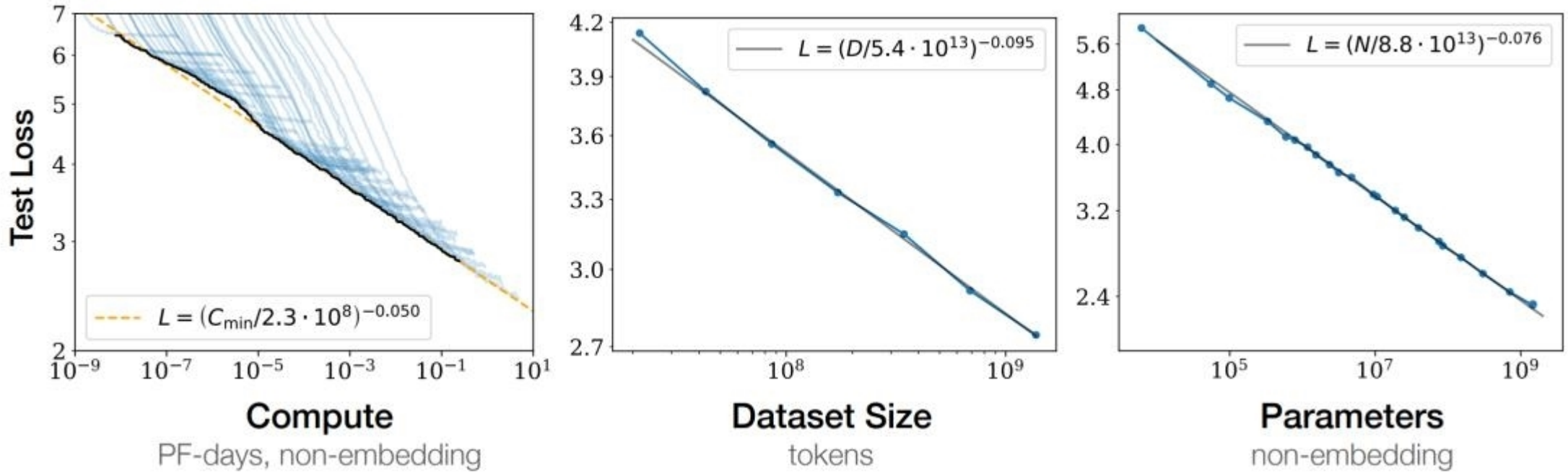
图：OpenAI决策路径

2015-2016	2017-2018	2018-2019	2018-2019	2018-2019	2019-2020	2020-2021
OpenAI 的决策						
早期 ML Engineering 能力和基础设施建设没有落后于行业，甚至目前比 Google 内部的还好用。	从 Unsupervised sentiment neuron 工作开始，逐渐将精力和关注点分配更多给语言模型上。	迅速和深度转向 Transformer，没有在 CNN/RNN 等上一代特征提取器上浪费时间。	在行业对强化学习的效果充满争议的情况下，在 Dota 及之后的项目中坚持探索深度强化学习。	在语言模型中坚持了仅有上文背景的 GPT 式生成式路线，没有追随 BERT 狂潮陷入理解式路线。	团队持续思考 Scaling Law 的问题，在 Transformer 基础上押注大规模数据和算力。	在长期强调安全和使用无监督强化学习的情况下，在 GPT-3 工作完成后迅速引入人类反馈。
当时市场的主流认知						
• AI 的突破是一项研究工作，而非工程问题； • 每个探索 AGI 的公司在工程能力和基建并不会有明显差距。	• OpenAI 的这个工作是优化别的任务时的副作用，歪打正着； • 语言模型不是通往 AGI 的道路。	• Transformer 彻底抛弃了之前 CNN、RNN 等网络结构； • 前几年统治 AI 进展的 CV 圈并不买账 Transformer。	• 深度强化学习的效率非常低； • 强化学习设置奖励函数非常 tricky； • 它会陷入局部最优，并且通常难以稳定复现效果。	• BERT 代表着未来，GPT 只是基于 Transformer 的过渡性技术； • GPT 白白丢掉了下文的信息，在许多自然语言理解任务上都难以和 BERT 竞争。	• AI 的进步来源于算法的创新； • 算力在过去 10 年的进步不一定在未来 10 年持续。	• 随着模型变得更智能，Alignment 问题可以自动解决，人类反馈多此一举； • 人类反馈违反了无监督的原教旨并且缺少可拓展性。

OpenAI的“暴力美学”：大算力和大数据

- OpenAI 在《Scaling Laws for Neural Language Models》中提出语言大模型所遵循的“规模法则”(Scaling Law)。
- Scaling Law 说明：通过独立延长模型训练时间(Compute)、增加训练数据量(Dataset Size)或者扩大模型参数规模(Parameters)，预训练模型在测试集上的 Test Loss 都会单调降低，从而使模型效果越来越好。
- 我们认为，在 Scaling Law 的框架下，只要追加数据与算力，大模型的能力就能持续增强。对于OpenAI而言， **目前大模型的最大限制是数据和算力的总量。**

图： Scaling Law：规模越大， 模型越精确



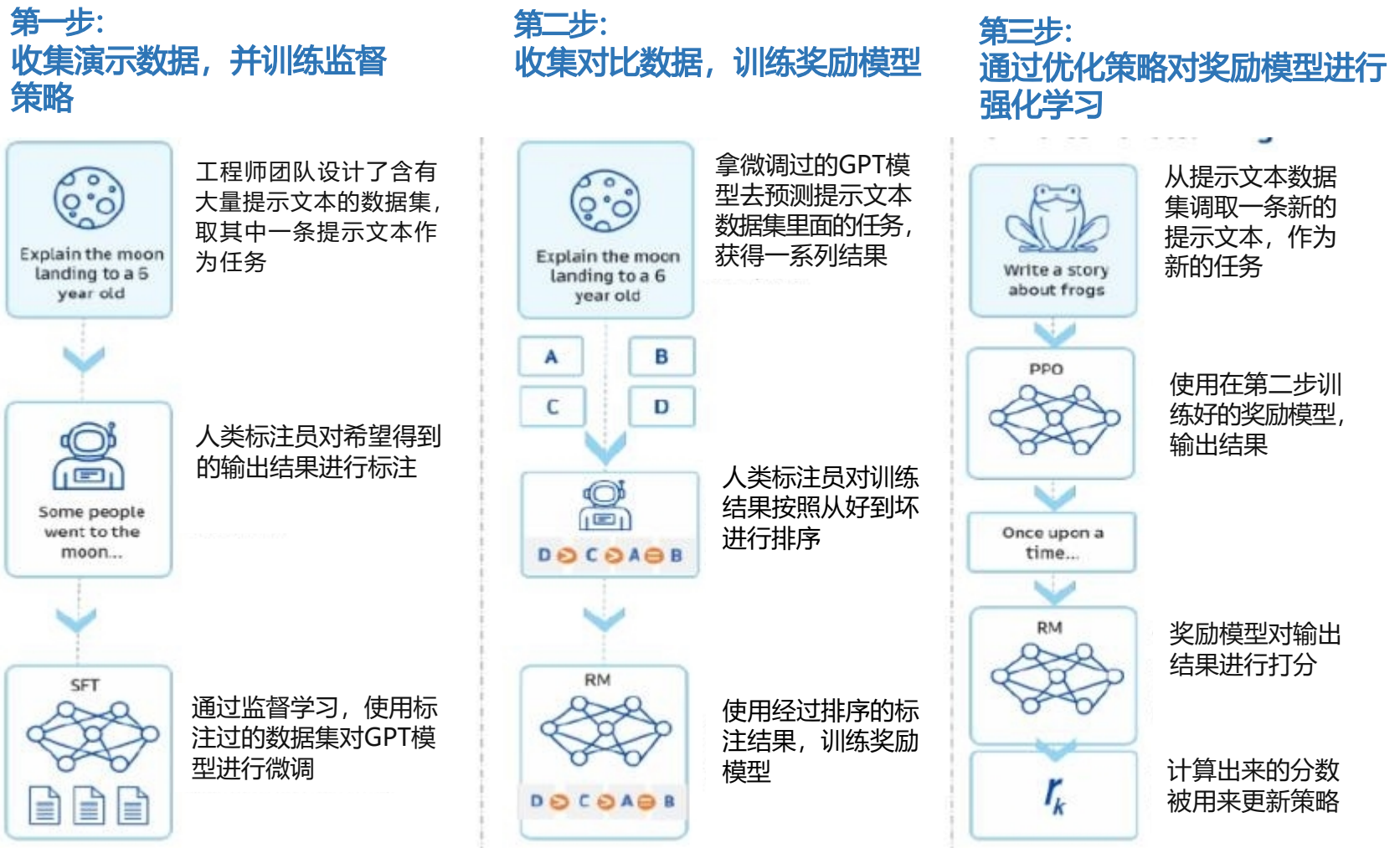
精细化策略+标注提升ChatGPT模型效果

- **预训练大模型分为上游(模型预训练) 和下游(模型微调)两个阶段。** 上游阶段主要是收集大量数据，并且训练超大规模的神经网络，以高效地存储和理解这些数据；而下游阶段则是在不同场景中， 利用相对较少的数据量和计算量，对模型进行微调，以达成特定的目的。

- **ChatGPT的训练过程也遵循预训练大模型的基本原理。** 结合了监督学习和强化学习，并且通过人工标注让模型更好地区别回复的好坏。

- 我们认为， ChatGPT在模型和数据等环节进行了大量的细节优化， 高质量的海量数据加上充分的训练，人工和算法的有机配合，使ChatGPT在模型层面实现领跑。

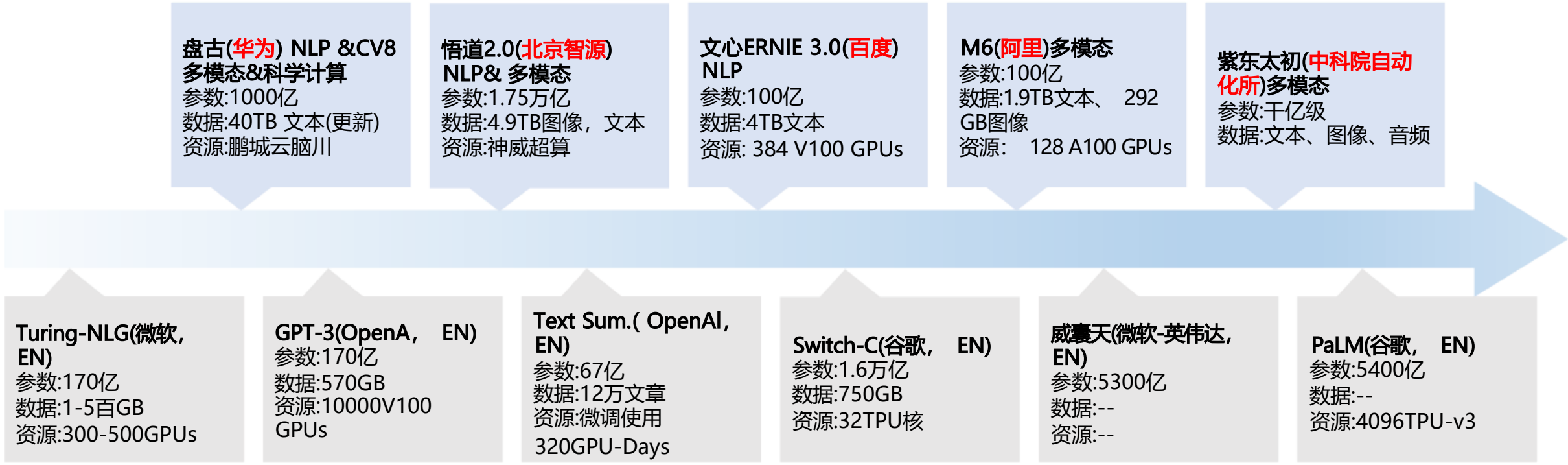
图： ChatGPT的训练原理



大模型开启新一轮大国博弈

- 预训练大模型是现阶段人工智能的集大成者， 代表了统计学习流派的最高成就。在新一代技术未出现前， 它将是人工智能研究和开发的最强武器。**围绕大模型的研发和落地，中美之间已经展开了新一轮的竞争。**
- 中国科学技术部高新技术司司长陈家昌，于2023年4月3日在国务院新闻办公室新闻发布会上表示， 在人工智能方面， **科技部专门加强顶层设计，成立人工智能规划推进办公室， 启动实施新一代人工智能重大科技项目。**
- 国内科技企业纷纷对ChatGPT发表看法，百度、华为、腾讯、阿里巴巴等大多数头部企业表示，已经拥有、在研对标ChatGPT相关的模型及产品。

图：中美大模型对比

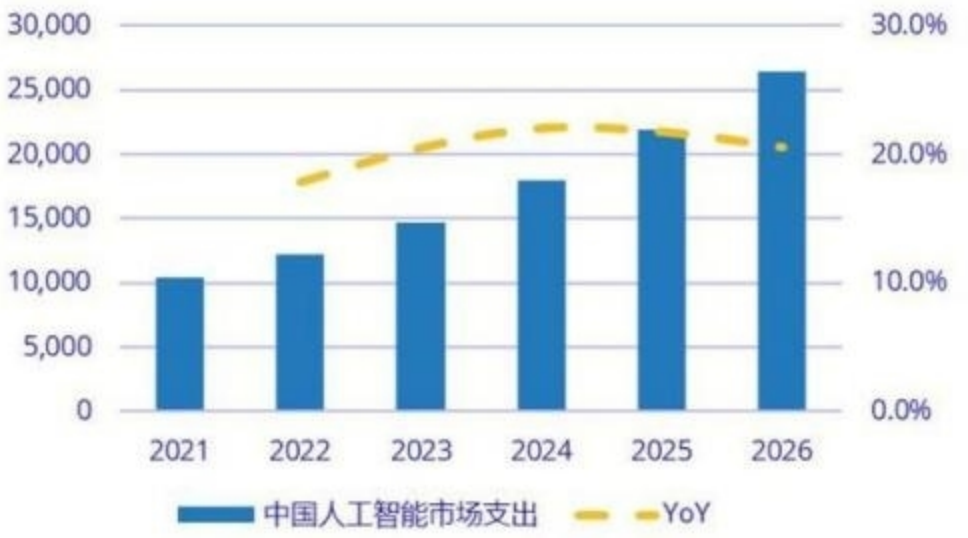


目录

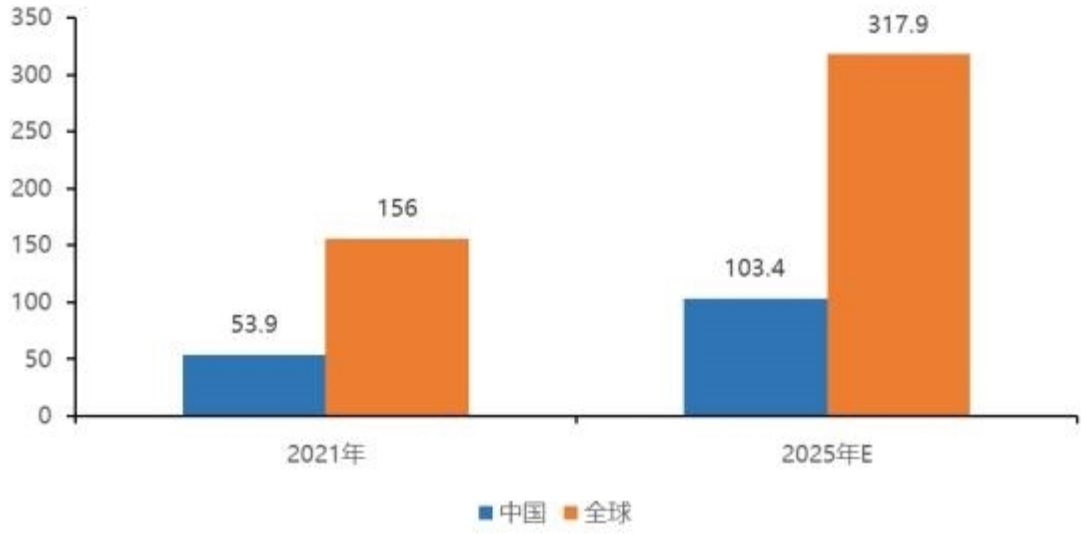
大国AI竞赛，国内AI支出规模有望高速增长

- 据IDC，中国人工智能(AI)市场支出规模将在2023年增至147.5亿美元， 约占全球总规模十分之一。 2021年中国加速服务器市场规模达到53.9亿美元(约350.3亿人民币)， 同比+68.6%； 预计到2025年将达到103.4亿美元。年复合增长率为19%，占全球整体服务器市场近三成。
- 我们认为，预训练大模型是现阶段人工智能的集大成者，代表了统计学习流派的最高成就。在新一代技术未出现前，它将是人工智能研究和开发的最强武器。围绕大模型的研发和落地，中美之间已经展开了新一轮的竞争。因此，国内人工智能的支出增速有望超过IDC的预测。

图：中国人工智能市场支出预测(百万美元)

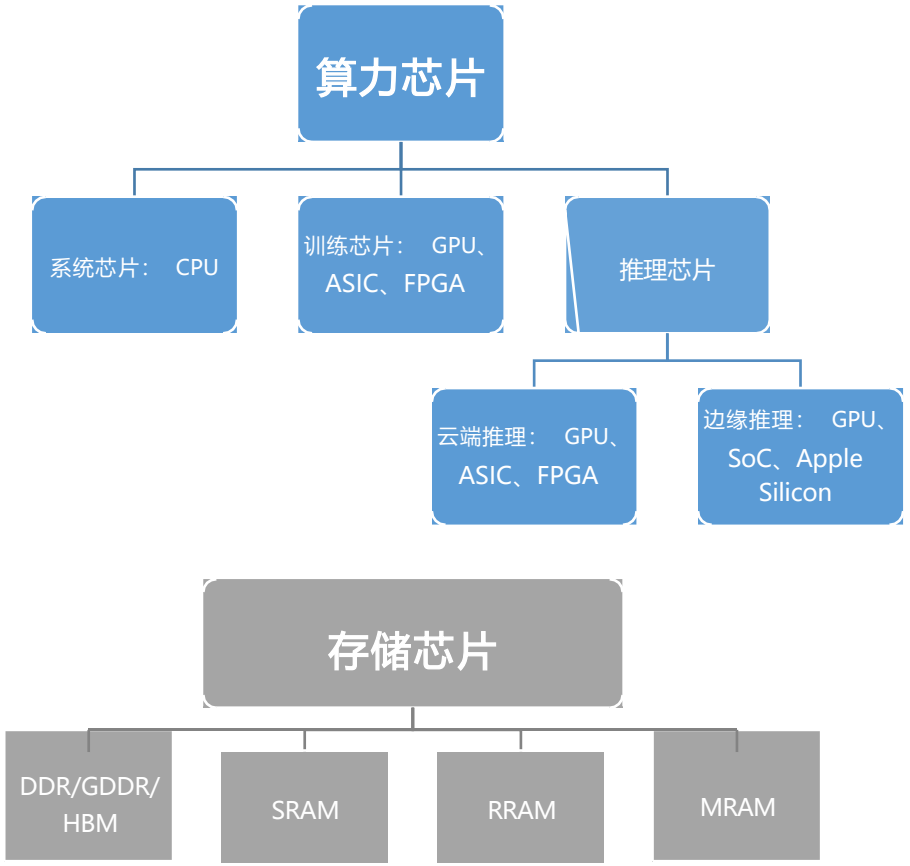


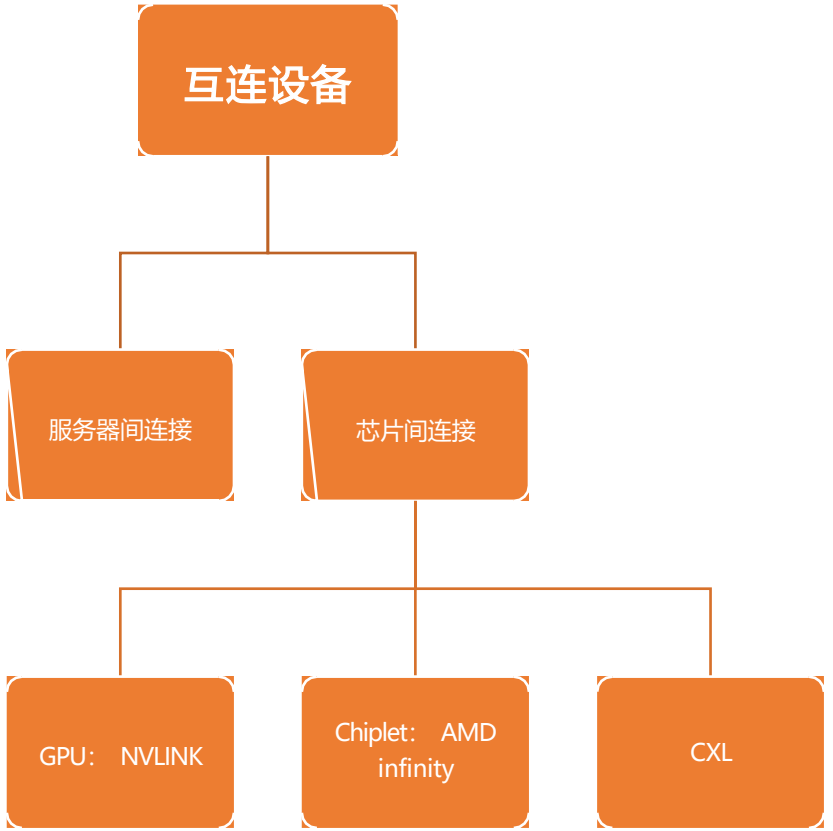
图：全球及中国AI服务器市场规模(亿美元)



算力芯片主导AI计算市场

- AI 分布式计算的市场主要由算力芯片 (55-75%)、内存 (10-20%) 和互联设备(10-20%)三部分组成。美国已限制对华销售最先进、使用最广泛的AI训练GPU—英伟达 A100以及H100，国产算力芯片距离英伟达最新产品存在较大差距， 但对信息颗粒度要求较低的推理运算能实现部分替代。
- **我们认为，训练芯片受限进一步强调了高制程芯片设计、代工的国产替代紧迫性。而随着人工智能的应用普及，推理芯片的市场需求将加速增长。**

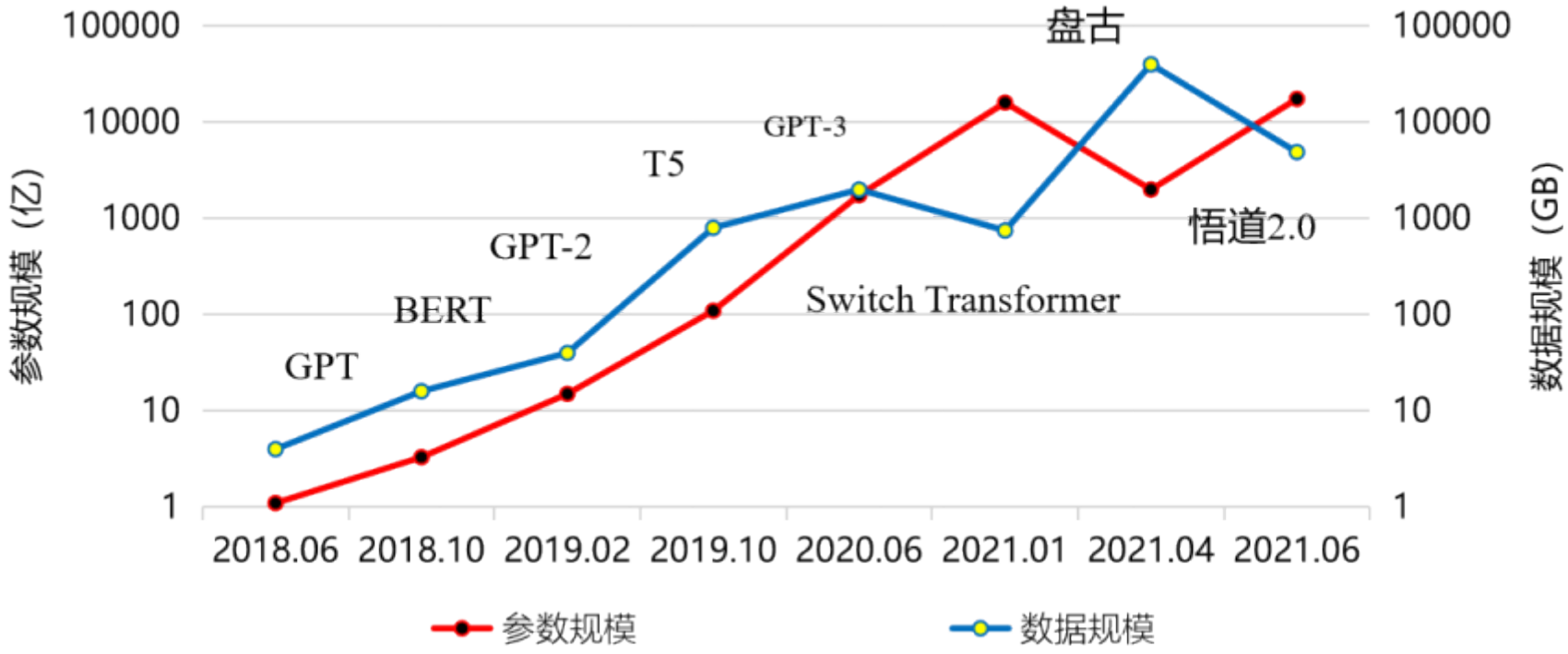




AI模型数据规模增长， AI算力需求井喷

- 当前，预训练模型参数数量、训练数据规模按照 300 倍/年的趋势增长，继续通过增大模型和增加训练数据仍是短期内演进方向。未来使用更多种图像编码、更多种语言、以及更多类型数据的预训练模型将会涌现。
- **当前算力距离AI应用存巨大鸿沟。 根据 Open AI 数据，模型计算量增长速度远超人工智能硬件算力增长速度，存在万倍差距。** 英特尔表示， 目前的计算、存储和网络基础设施远不足以实现元宇宙愿景，而要想实现真正的元宇宙，目前的计算能力需量要再提高1000倍。

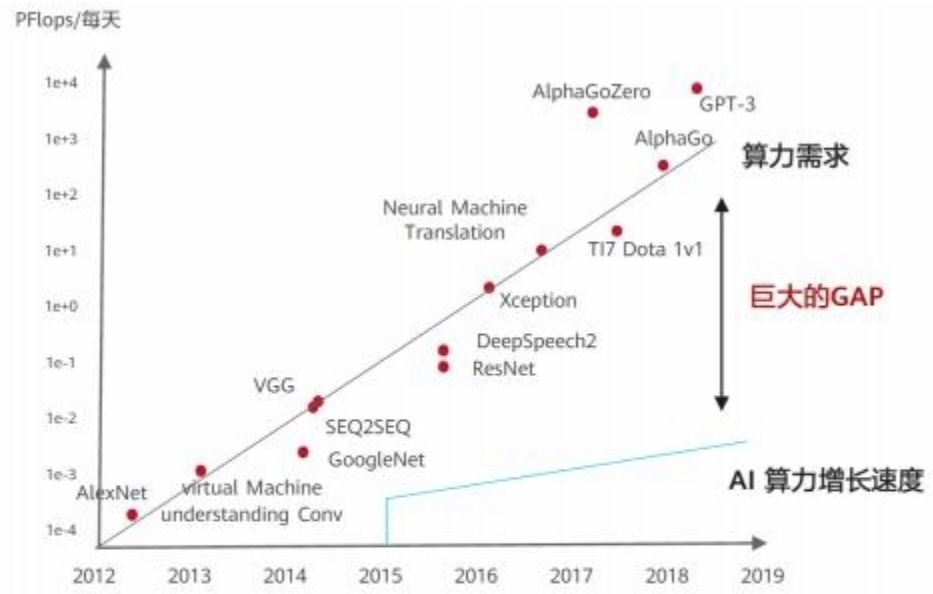
图：大模型参数量和训练数据规模增长迅速



AI模型数据规模增长， AI算力需求井喷

- 据IDC预计， 2021-2026年期间，中国智能算力规模年复合增长率达52.3%。2022年智能算力规模将达到268.0 EFLOPS，预计到2026年智能算力规模将进入每秒十万亿亿次浮点计算(ZFLOPS）级别， 达到1,271.4 EFLOPS。
- **运算数据规模的增长，带动了对AI训练芯片单点算力提升的需求，并对数据传输速度提出了更高的要求。**

图： 2012至2019年算力需求增长近30万倍



AI模型训练花费不菲

GPT-3	460万美金	10000块GPU * 13天

图：中国智能算力规模百亿亿次浮点运算/秒(EFLOPS)