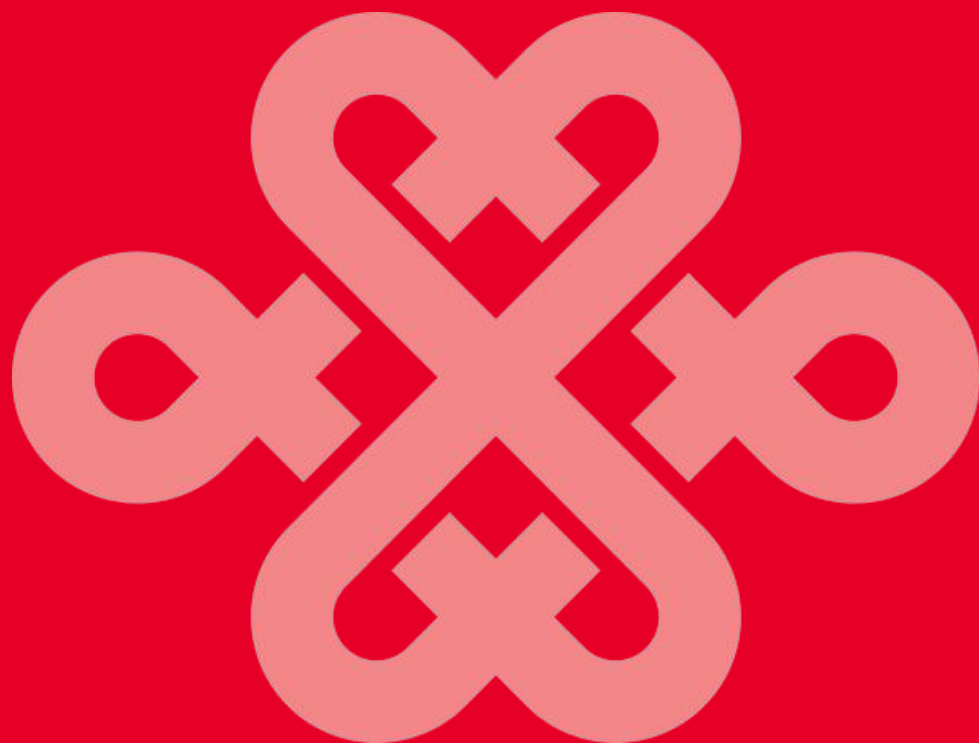




探索算力时代的信创化体系

联通（北京）产业互联网有限公司

2023年11月

一. 算力时代到来，聚焦信创化体系

二. 国产硬件革新，对标英伟达算力

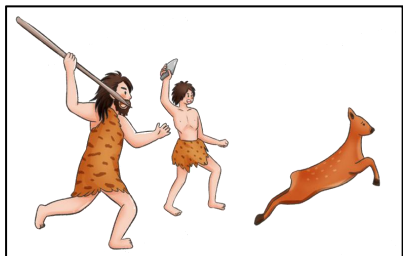
三. 模型软件齐放，适配信创化平台

四. 行业模型并进，toB & toC共生

五. 安全体系保障，信创化健康发展



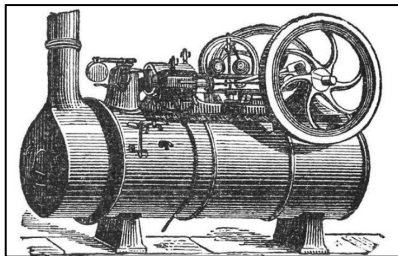
算力时代：人类发展史的又一次变革



人力时代



畜力时代



煤力时代



电力时代



算力时代

算力时代三大定律：

- 第一定律：**时代定律** 算力就是生产力

- 人类生产力的发展已经从原始经济的人力时代、农业经济的畜力时代、工业经济的电力时代，来到数字经济的算力时代。

- 第二定律：**增长定律** 算力每12个月增长一倍

- 受底层技术融合发展及加速成熟，智能算力增速迅猛。随着元宇宙技术成熟与应用场景拓展等影响，算力需求的资源进一步增大，推动算力模式呈现爆发式增长态势。

- 第三定律：**经济定律** 算力每投入1元，带动3-4元GDP经济增长

- 全球各国算力规模与经济发展水平，呈现出显著的正相关关系。算力在驱动社会和产业发生深刻变革的同时，也将产生显著的经济价值。

“坚持创新驱动。创新是发展的强大动力。我们要顺应科技发展趋势，以更加积极姿态推动科技交流合作，携手打造**开放、公平、公正、非歧视**的科技发展环境。要加速数字化转型，缩小数字鸿沟，加快落实《亚太经合组织互联网和数字经济路线图》，支持**大数据、云计算、人工智能、量子计算**等新技术应用，不断塑造亚太发展新动能新优势。”

习总书记在APEC第三十次领导人非正式会议上的讲话

需要破解的问题：

英伟达租卖化
大模型套壳化
新功能模仿化

算力时代：人类发展史的又一次变革

算力时代1.0: **算力** + **算法** + **用户** + **场景**
AI处理器 大语言模型 2C/2G/2B 科研/政务/智慧社区/传媒/医疗...

算力时代2.0: **算网** + **生态** + **制度**
算力+算法+网络 供需+消费+创新 安全+围栏+伦理

算力时代3.0: **生产力** + **创造力** + **约束力**
软硬件人工替代 全方位科技加速 高难度安全压力

经济和技术的角逐

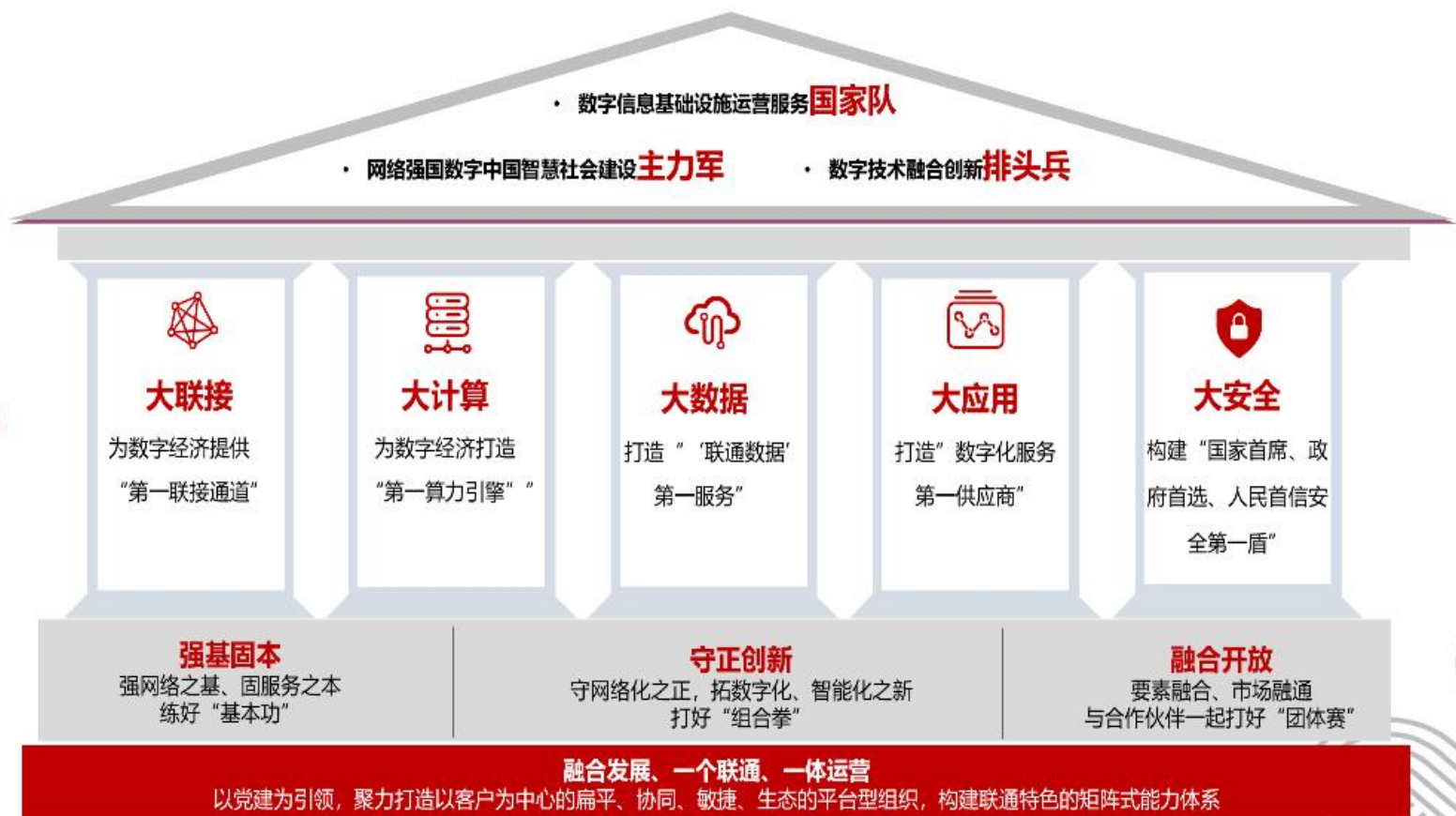
- 2021年4月：美国政府进一步扩大禁令范围，**限制**英伟达向中国出口 **A100 和 H100** 等先进 AI 芯片
- 2022年10月：美国将**31家**中国公司、研究机构和其他团体列入 “**未经核实的名单（UVL）**”
- 2023年10月：美政府对华高端芯片发展的打压**再度升级**。英伟达**A800和H800**在内的芯片对华出口将受到影响
- 2023年10月25日：美国政府针对适用于总处理性能大于等于4800并为数据中心设计或销售的产品的**相关出口管制立即生效**，影响公司**A100、A800、H100、H800和L40S**产品。
- 中国对美国的芯片制裁采取了多重反制措施，包括修订《中国禁止出口和限制出口技术目录》等，限制关键技术的出口，同时加强国内创新和自主研发。

- **芯片产业领先地位**
- **前沿领域技术创新**
- **芯片产业链完整**



- **消费市场规模巨大**
- **自主研发能力显著提升**
- **政府政策强力支持**

- 中国联通将以“1+9+3”新战略规划体系为指引，通过构造**算力网络四位一体**的体系架构，打造**算力网络四位一体核心能力**，创造**算力网络技术与产业新范式**，“三箭齐发”推进算力网络建设。
- 中国联通战略规划纲要指出，**建立京津冀协同高效的算力中心体系**，推动形成**4000PFLOPS**总算力规模的人工智能算力基础设施；铸就发展引擎，形成算网融合差异化优势，为数字经济打造“**第一算力引擎**”，2025年，实现总算力超过**1EFLOPS**。



《中国联通算网融合发展行动计划》

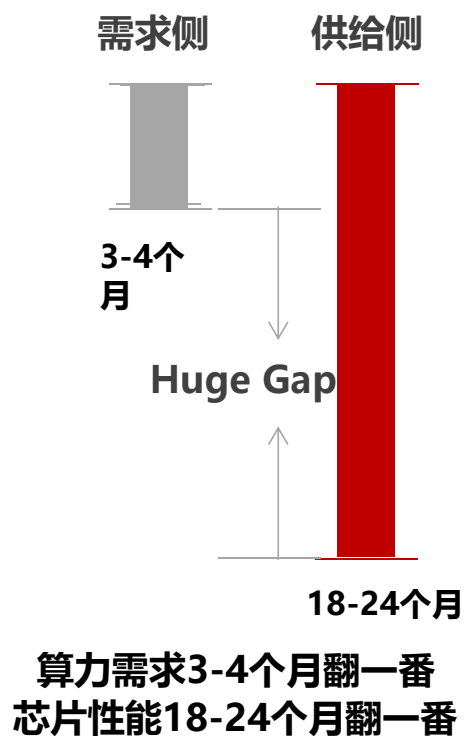
- **坚持规划引领**
布局全国一体化新型数据中心资源
- **坚持市场驱动**
打造全国一体化算网融合产品体系
- **坚持创新驱动**
打造全国一体化算力网络差异化优势
- **坚持智慧运营**
建立全国一体化算网运营支撑体系
- **坚持开放合作**
深化全国一体化算网融合新生态

信创之路：打造网络强国“算力+算法”信创优势

- 算力正在成为影响国家综合实力和经济发展的关键性要素，**算力建设已然成为国家高质量发展的战略级方针。**
- 中国市场的庞大需求，美国芯片断供，为国产AI芯片的崛起带来难得机遇。加强自主研发、提高产业竞争力、**加快芯片国产化成为我国芯片产业发展的关键。**

建强国产化算力

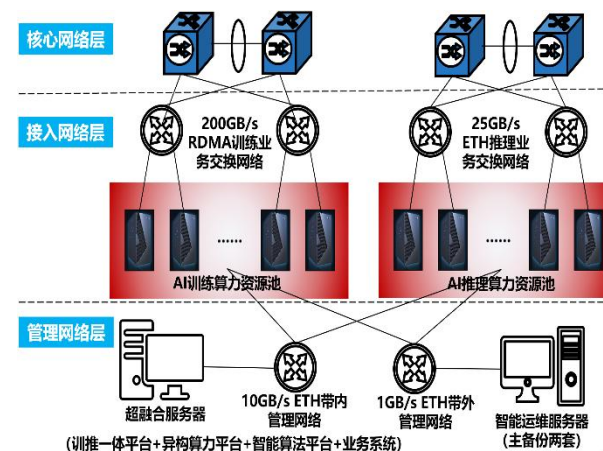
- 顺应大模型趋势，**算力需求急剧攀升；**
- **AI芯片实现自主性供给。**当前国内寒武纪、华为、海光、昆仑芯、燧原等厂商成熟度较高。
- 截至2023年2月，我国目前已投入运营和在建的**人工智能计算中心达23个；**



需要快速增加算力规模 ←

建优国产化算网

- 算网是“**主动脉**”，实现**算力、数据的高效承载与流通；**
- “**大带宽、低时延、高可靠、敏捷拆建**”是大模型对算网的要求。

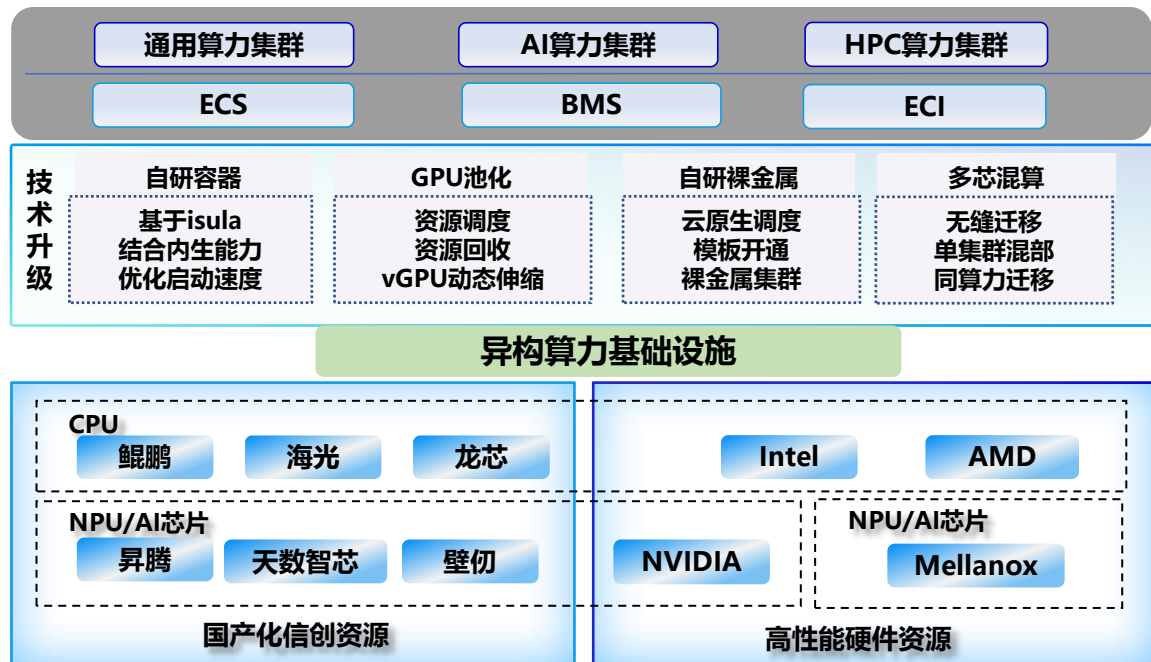


- 推进网络设施与算力设施配套部署
- 建立智算集群间的高速网络及互联平台，**强化算力资源统筹调度，提高算力利用效能**

- 中国联通建成数据中心900余个，机架总规模超过36万架，承载服务器超百万，形成“5+4+31+X”资源布局，建设高算力、高安全、绿色低碳新型数据中心，实现大规模算力部署，提升算力服务品质和利用效率，支撑不同维度、不同层次全面多样的数字化需求。
- 智能化安全可信行业云：打造了行业算力专区，上线高性能算力，叠加AI、GPU等智能组件。基于联通云7信创云版本，适配超95%的国产化主流软硬件产品，支持X86+ARM+OpenPower混部，多芯协同，无缝切换，屏蔽底层差异、快速实现基础软硬件替代，实现一云多芯、全栈产品自主开发，提供完善的“信创+安全+密码”服务能力。



联通云多样性算力供给架构



■ 自研容器 + DevSecOps

➢ 容器引擎全链路安全可控，部署效率提升50%

■ 自研容器 + 自研信创裸金属

➢ 解决“卡脖子”问题，100%兼容OCI\CRI

■ 无缝迁移

➢ 多技术路线芯片混部，应用无感无缝迁移，更加安全可信

■ 多场景适配，算力解耦

➢ 异构算力混合编排框架，提供产品化供给的多样算力

一. 算力时代到来，聚焦信创化体系

二. 国产硬件革新，对标英伟达算力

三. 模型软件齐放，适配信创化平台

四. 行业模型并进，toB & toC共生

五. 安全体系保障，信创化健康发展





NVIDIA (GPU)

- **Tesla系列**：专业级显卡。主要用于深度学习、人工智能和高性能计算。如 **A100 (A800)**、**H100 (H800)**、**A30**、**A40**、**V100**、**P100**...
其中，A800/H800是针对中国特供版（低配版），相对于A100/H100，主要区别：
A100的Nvlink最大总网络带宽为600GB/s，而A800的Nvlink最大总网络带宽为400GB/s。
H100的Nvlink最大总网络带宽为900GB/s，而H800的Nvlink最大总网络带宽为400GB/s。
- **GeForce系列**：消费级显卡，主要应用于游戏娱乐领域，如**RTX 3090**、**RTX 4090** ...
- **Quadro系列**：专业级显卡，主要用于专业可视化设计，如**RTX 6000**、**RTX 8000** ...

其他国外AI处理器（加速卡）

- **AMD**：GPU MI300X，采用CDNA 3架构，拥有8个XCD核心，最多304个CU。
- **Intel**：Xeon Phi，是首款英特尔集成众核（Many Integrated Core, MIC）架构产品。
- **Google**：TPU，针对服务器端AI模型推理，主要用于Google云计算和数据中心。



国产AI处理器（加速卡）

- **华为**：昇腾910（用于训练和推理），昇腾310（用于推理），采用自家设计的达芬奇架构。
- **阿里**：含光800，基于RISC-V和阿里自有算法。
- **寒武纪**：思元370、思元590，采用新一代智能处理器架构 MLUarch03。
- **百度**：昆仑芯，采用的是其自研XPU架构。
- **曙光**：DCU，拥有AMD的x86技术授权，以GPGPU架构为基础。
- **燧原**：云燧T1x/T2x，i1x/i2x，采用全自研的GCU-CARA全域计算架构。



国产算力：国产化需求加大，昇腾较为成熟（1/2）



- 1、算力国产化趋势明显：美国限制高端芯片对中国出口，2023年10月25日，美国政府针对英伟达A100、A800、H100、H800和L40S等产品进行管制，算力国产化趋势进一步加大。
- 2、国产算力昇腾最优：国产化AI芯片企业中，华为昇腾技术成熟度最高，国产市场占比第一。华为从18年推出昇腾910A芯片以来，华为在昇腾上每年投入15亿+美金，持续构建昇腾生态。目前，80%+行业TOP伙伴适配昇腾，采用昇腾算力平台，其次，在出口管制条件下，英伟达进一步推出阉割版的产品，性能下降、价格增高，华为昇腾芯片的优势将进一步提高，加快生态完善。其中，相对领先的寒武纪思元系列存在底层软件能力缺失、大规模生产限制等问题，短期难以大规模国产化部署。

内容	华为	寒武纪	阿里含光	曙光	燧原	百度
拓展路径	基于“一中心四平台”思路建设新型基础设施，推动“政、产、学、研、用”的闭环	专注芯片及基础软件平台研发，通过行业热点,如自动驾驶引导建设	自研架构，通过软硬件协同设计实现性能突破。基于RISC-V和阿里自有算法，可以获得高吞吐量和低延迟的高性能体验	推出5A级智算中心解决方案，打造融合新型智能算力基础设施	提供训推一体集群，用于智慧城市、科研院所及企事业单位算力底座	采用云智一体战略定位，打造全栈自研的AI基础设施“AI大底座”，使用大模型赋能行业
规划路标	训练主要使用昇腾910芯片：已实现大规模商用，提供从迁移、适配、开发、训练的完整技术支撑	MLU(机器学习处理器)已发布两代产品，第三代已回片： 2020年：MLU 290 2022年：MLU 370 2023年：MLU 590	2019年发布数据中心AI推理芯片含光800，性能达78563 IPS，能效达500 IPS，峰值算力达820 TOPS的高性能人工智能推理芯片	DCU(深度计算处理器)使用AMD GPGPU架构授权，已发布两代：DCU一代已实现商业化应用DCU二代已发布	已发布两代产品： AI训练加速卡云燧T1x/T2x AI推理加速卡云燧i1x/i2x	已发布两代，主要用于推理： 2018年：昆仑芯一代 2021年：昆仑芯二代 2024年：昆仑芯三代
算力类型	训练、推理	训练、推理	推理	主打HPC，具备训练/推理能力	训练、推理	推理
可交付性	采用全栈自研技术方案，经大规模商用建议，可稳定交付	主力交付MLU 370和290；MLU590可小规模交付（台州200片，提供60P@FP16）	含光800已成功应用在数据中心、边缘服务器等场景	少量训练算力由DCU提供，其余算力采用寒武纪或者英伟达	尚未有大规模商用案例，可交付性未知	训练算力通常由英伟达提供，昆仑芯提供推理算力
产业布局	已上线25个局点，总算力规模5000P+，服务3000+家企业，高校和科研机构，运营效果良好	1. 浙东南数字经济产业园 2. 沈阳汽车智算 3. 南京智算中心 4. 珠海横琴	含光800已成功覆盖阿里巴巴丰富的应用场景；部署于城市大脑、拍立淘，搜索和广告推荐等多个业务	1、长沙人工智能计算中心（模块二） 2、合肥先进计算中心（一期）	之江实验室：采用云燧一代AI加速卡，提供FP32算力，结合之江的国产自研天枢人工智能开源平台，支撑科研。	行业场景：智能制造（吉利汽车）、智慧城市（海淀城市大脑）、金融（邮政储蓄银行） 政府场景：22年底发布AI底座，面向政府拓展

- **3、华为昇腾生态完善：**中国大模型近 **1/2 选择昇腾**（其余选择英伟达），在**国产生态中处于领先地位**，昇腾已适配GLM、LLAMA、GPT等10+业界大模型，可提供全流程工具辅助大模型迁移，助力业务快速上线。其深度开放，分层解耦的特点，全面支持业界各类框架、加速库及三方社区生态，可提供大规模并行训练、工程化调优原厂支持。对于常见大模型，华为提供便捷的迁移适配工具，**一般在一个工作日可以完成**；对其他场景，有专门团队做针对性的迁转适配，未来迁移难度进一步降低。

在解决迁移适配成本、开发效率、技术门槛方面，华为昇腾910芯片具有丰富生态优势

为了降低开发门槛、减少适配成本、提升迁移效率，昇腾算力集群提供一站式AI平台，覆盖机器学习/深度学习AI开发全流程，通过多元网络自动搜索，大大降低模型人工调参的成本和门槛，优势主要表现在：

- **便捷的算法训练：**昇腾提供强大的基础加速库、丰富的模型开发套件，以及易用的开发体系，让训练更高效，创新更便捷、开发更简单。
- **亲和的算子开发：**推出原生支持C/C++的开发语言，降低算子开发门槛，同时提供昇腾硬件亲和的数据结构，保持算子性能。
- **高效的模型开发：**华为提供从迁移、开发到调优的整套工具和套件。支持业界主流的10多个模型套件及动静态图统一，兼顾开发及执行效率。
- **丰富的应用开发：**华为提供SDK，把一类场景的开发经验和行业知识积淀下来，通过插件化像搭积木一样编排应用，以此大幅度降低门槛。
- **高性能的大模型开发：**华为推出大模型开发使能平台，让大模型更易开发、易适配、易部署。提供高性能API封装，简化大模型的开发难度。
- **轻量化的大模型部署：**采用昇腾平台可以实现20倍的大模型压缩，大幅降低大模型的推理成本。

- **4、践行国家新兴战略：**按照**国资委“9+6”战略新兴产业布局**的总体规划，不断的提高核心竞争能力，增强核心的功能，促进战略性的新兴产业的壮大，加快壮大新一代**信息技术、生物技术、新能源、新材料、高端装备、新能源汽车、绿色环保以及航空航天、海洋装备九大产业**，聚焦**算力网络、人工智能、6G、大数据、安全、能力中台**六个领域承担新的战略任务。推动互联网、大数据、人工智能等同各产业深度融合，推动先进制造业集群发展，构建一批各具特色、优势互补、结构合理的战略性新兴产业增长引擎，**培育新技术、新产品、新业态、新模式**。北京市昇腾人工智能计算中心作为首例运营商大规模部署昇腾智算生态的标杆性项目，具有**科技创新，产业控制，安全支撑**重要作用。

- **初代昇腾910A产品：**2018年推出，产品研发初衷对标英伟达A100，但在计算能力上有一些能力缺失，客户诟病比较多。距离A100差距较大。
- **新一代昇腾910B产品：**2023年7月上市，研发时间5年，910B相对910A的关键升级（1）FP32单精度浮点运算能力提升；（2）芯片间互联带宽从100G提升到200G；这两个能力升级，对计算能力和运行效率有大幅提升，具有跨时代意义，**性能已经基本达到英伟达的水平**，可满足企业智算训练服务需求。
- **昇腾下一步研发方向：**新产品预计在2025年以后推出，持续提升芯片计算能力与效率，持续提升产品生态能力。
- **昇腾支持混合精度训练：**昇腾910支持FP32+FP16混合精度训练模式，完全满足业界客户对于AI训练的需求。FP64虽然精度更高，但不适合做AI模型训练，使用FP64精度通常用于高性能计算领域，AI领域数据特征信息具有一定的冗余性，一般不需要使用FP64高精度，且同等算力下FP64的设备数量更多、能耗也高，导致用户使用FP64成本更高，因此FP64不适合做AI模型训练。

一. 算力时代到来，聚焦信创化体系

二. 国产硬件革新，对标英伟达算力

三. 模型软件齐放，适配信创化平台

四. 行业模型并进，toB & toC共生

五. 安全体系保障，信创化健康发展



- 国内信创软硬件厂商在兼容适配性、性价比、集成场景的深度与广度，以及行业know-how等方面，已然向前迈进了一大步在基础软件行业，国内桌面端操作系统行业龙头**统信UOS**，其所服务的客户则覆盖了超**70%**的党政机关、**45%+**的央国企、**65%+**的部委，以及**85%+**的金融行业企业。

国内外信创行业对比

中标麒麟

华为鲲鹏

深桑达

海量数据

纳斯达

金山办公

华大九天

华为昇腾

VS

Windows

Intel

亚马逊

Oracle

惠普

Office

新思科技

英伟达

国产化痛点分析

- 企业业务场景日趋复杂，信创厂商面临挑战。随着业务的快速增长，多源异构的海量数据，难以用数据量化的复杂场景，都是摆在信创厂商面前的难题。
- 大量信创方案仍在测试和形成过程中，产品能力有限。产品能力不足，直接导致信创整体进程受限。
- 国内软硬件行业人才储备不足；同时，对于人才的投入不足，且短期内投入难以形成价值转化。
- 国内信创生态较弱，存在应用范围较窄、兼容性差、性能无法满足要求、通用性低等问题。

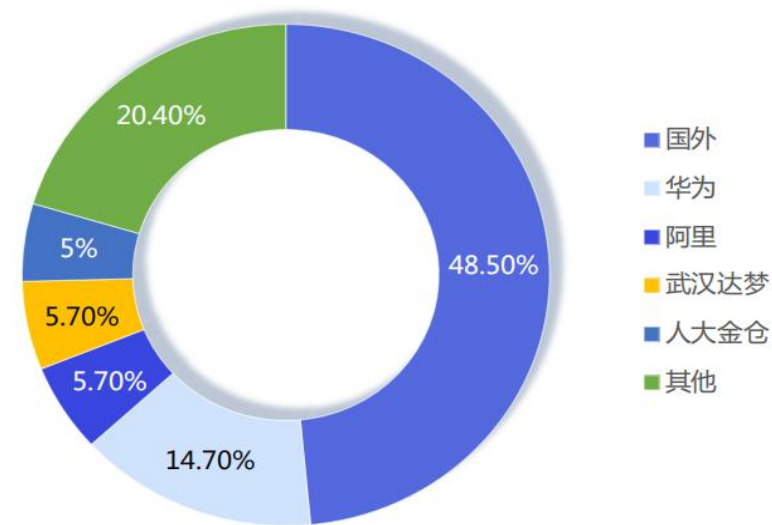
- 信创行业正被“机遇”和“挑战”两股相反的力量同时拉扯。一端是信创生态正快速裂变，市场招投标数据大幅增长，多个市场进入爆发前夜。另一端，来自人才、生态、产品端的现实挑战。目前国内信创行业在技术和市场格局等已出现两大明确趋势：
 - 第一，未来3-5年，国内信创行业将进入并购整合期，形成巨头为主导、少数企业并存，共同替换海外厂商市场份额的格局。
 - 第二，AIGC将进一步助推信创产品在各垂直行业的深入应用。

部分国产数据库厂商融资案例

厂商名称	融资轮次	最新融资时间	部分投资方
人大金仓	战略融资	2021.12	太极股份、中国电科、北京高新创投、勤益投资
达梦数据库	战略融资	2019.12	丰年资本、信风投资、中电智慧基金、中国软件
南大通用	D轮	2022.10	君联资本、耀途资本、国投创合、苏州国发创投、歌斐资产、大唐电信、东华软件、拓尔思
PingCAP	E轮	2021.07	红杉中国、新加坡政府投资公司(GIC)、GGV纪源资本、经纬创投、Access Technology Ventures
偶数科技	B+轮	2021.08	腾讯投资、红杉资本、红点中国、金山云
Zilliz	B+轮	2022.08	高瓴创投、五源资本、云启资本、靖亚资本、松禾资本

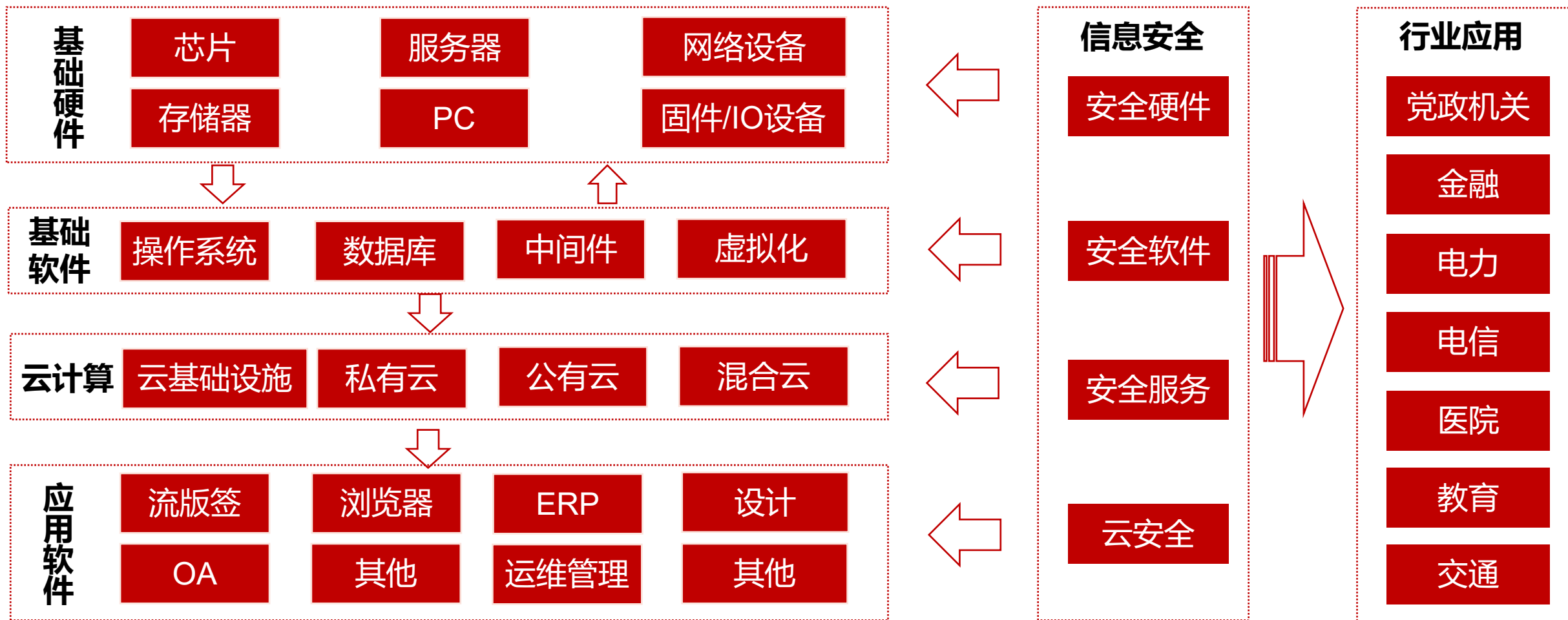
数据来源：天眼查，第一新声研究部

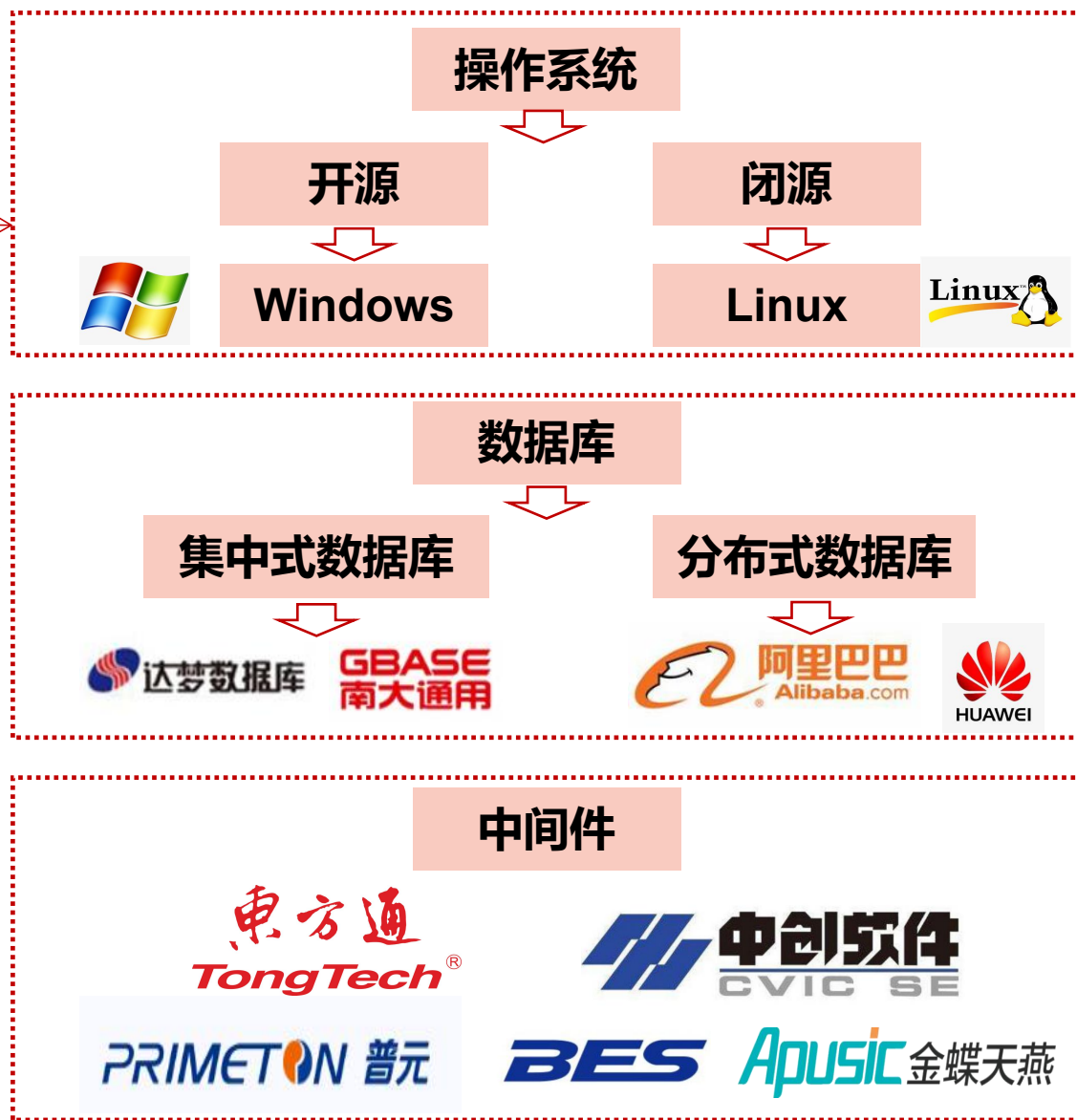
2021年中国数据库市场格局



数据来源：华经产业研究，第一新声研究部

- 我国信创行业已基本建立起从上游芯片到下游应用的替代产业链条。从产业链角度看，信创生态体系主要由基础硬件、基础软件、应用软件、信息安全及云计算平台几部分组成，其中**芯片、服务器、操作系统、数据库、中间件**是最重要的产业链环节。





- 我国在**操作系统**方面也取得了一定的进展。如在金融市场以**麒麟**、**统信UOS**、**红旗**等为代表的国产操作系统:在桌面操作系统领域表现突出的**统信UOS**;国产手机操作系统的杰出代表**华为鸿蒙**。
- **数据库**的服务市场主要集中在金融、电信、政府、制造、交通五个行业**人大金仓**、**达梦数据库**、**南大通用**等一批拥有高校背景的国产厂商成立，打破了甲骨文和国际商业机器的垄断格局。
- 国产中间件行业也在日益成熟，新基建背景下，将加速中间件国产化进程。国产中间件产品逐渐得到用户认可，在部分领域中已完全具备替代国外产品的能力，未来中间件产品的国产化替代步伐将持续加快国产中间件厂商将得到前所未有的发展机遇。

■ 在信创领域，麒麟操作系统目前占据主导地位，而以深度操作系统为基础开发的UOS也已快速崛起，并成为最受欢迎的民企国产操作系统，双足鼎立的市场格局已基本形成。

	麒麟	统信	普华	中科方德	一铭
产品类型	桌面OS、服务器OS	桌面OS、服务器OS	桌面OS、服务器OS	桌面OS、服务器OS	桌面OS、服务器OS
内核	Linux	Linux	Linux	Linux	Linux
核心股东	中国电子CEC	诚迈科技/统信	中国电科CETC/太极股份	中科院软件所	一铭软件
适配CPU	目前所有国产CPU	目前所有国产CPU	X86.Open Power. 龙芯、申威	兆芯、主流X86平台	Intel X86、龙芯
优势	中国Linux市场占有率第一，党政、国防办公领域占国产Os 90%以上市场份额	基于Deepin操作系统打造，最受欢迎的民企国产操作系统，深度绑定华为	中国电科CETC旗下企业	国家级项目技术研究中心	中国开源软件推进联盟副主席；“新三板”挂牌上市

数据库

- 根据国产数据库划分，从市场占有率划分达梦、人大金仓，从生态体系达梦电子系、人大金仓电科、海量华为生态，神州航天系，瀚高山东政府等，已有众多项目开始采用国产数据库，逐渐被市场所认可，在分布式数据库等新兴细分数据库市场，国产数据库表现尤为突出。

序号	数据库品牌	技术来源	体系
1	达梦	oracle	中国电子
2	海量数据库	OPENGNSS	华为生态
3	人大金仓	PosgreSQL	中国电科
4	南大通用	Informix	天津南开
5	神州通用	PosgreSQL	航天系

中间件

- 中间件是沟通操作系统、数据库与上层应用系统的桥梁。

国内外中间件对比

总排名	名称	分类	得分	公司/组织
2	Sharding-JDBC	数据库中间件	154.58	Apache软件基金会
3	ShardingSphere	数据库中间件	143.81	Apache软件基金会
9	Canal	数据库中间件	43.22	阿里云
11	SphereEx	数据库中间件	22.43	北京思斐软件技术有限公司
13	Sharding-Proxy	数据库中间件	2.3	Apache软件基金会
14	Sharding-Sidecar	数据库中间件	1.31	Apache软件基金会
15	DBLE	数据库中间件	0.8	北京爱可生信息技术股份有限公司
16	Cobar	数据库中间件	0.68	阿里云
17	轻舟	数据库中间件	0.62	杭州网易数帆科技有限公司
23	信也DAS	数据库中间件	0.24	信也科技
25	京东云JCQ	数据库中间件	0.08	北京京东尚科信息技术有限公司
33	zdal	数据库中间件	0.01	阿里云
35	TongRDS	数据库中间件	0	北京东方通科技股份有限公司
43	华为云DDM	数据库中间件	0	华为云
45	银信分布式数据处理	数据库中间件	0	北京银信长远科技股份有限公司

全产业链信创化替代：AIGC算力产业

AIGC产业算力服务商图谱

云服务



大模型



智算中心



各大智算中心

服务器



计算芯片



- 智谱AI提出了 GLM算法框架。该框架实则实现了BERT、GPT优劣互补，“既能在理解的同时，还能完成续写和填空”。在此框架之下陆续研发出了GLM-130B、ChatGLM-6B、ChatGLM2-6B等GLM系列大模型。实验数据显示，GLM系列大模型在语言理解精度、推理速度、内存占比和大模型适配应用方面都优于GPT。

大模型	基础架构	训练方式	量化	加速	跨平台能力
GPT	GPT	自监督预训练	——	——	NVIDIA
GLM	GLM	自监督预训练 多任务预训练	INT8/INT4	Faster Transformer	NVIDIA 海光DCU 昇腾910 申威
GLM优势	高精度: Big-bench-lite: +5.2% LAMBADA: +2.3% CLUE: +24.3% FewCLUE: +12.8%		普惠推理: 节省75%内存 可单台3090(4) 或单台2080(8) 进行无损推理	高速推理: EtPytorch 提速7-84倍 Megatron 提速25倍	跨平台: 支持更多不同的大规模 语言模型的适配和应用

一. 算力时代到来，聚焦信创化体系

二. 国产硬件革新，对标英伟达算力

三. 模型软件齐放，适配信创化平台

四. 行业模型并进，toB & toC共生

五. 安全体系保障，信创化健康发展

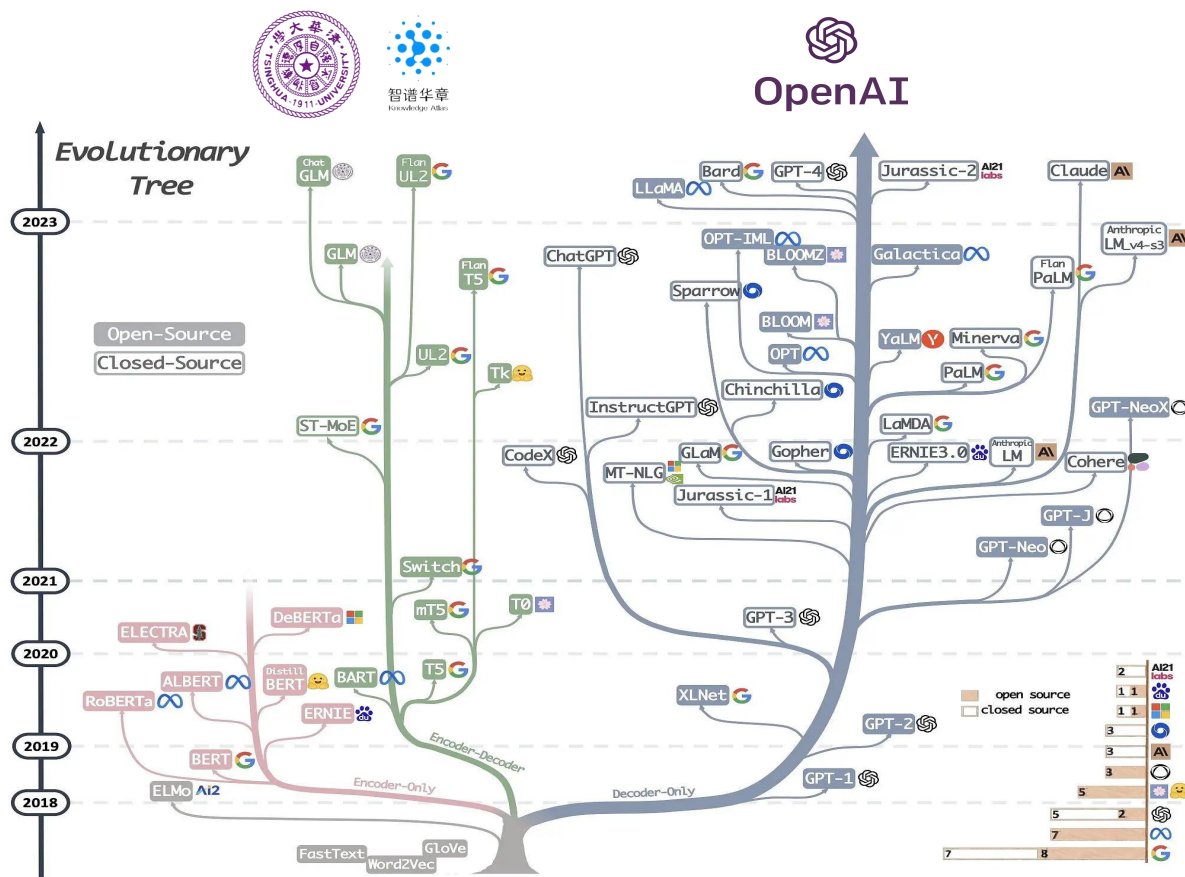


算法类型：全球LLM(大语言模型)体系发展分析

在 Transformer 模型中，**编码器负责理解和提取输入文本中的相关信息**。这个过程通常涉及到处理文本的序列化形式，例如单词或字符，并且用**自注意力机制（Self-Attention）**来理解文本中的上下文关系。

编码器的输出是输入文本的连续表示，通常称为嵌入（Embedding）。这种嵌入包含了编码器从文本中提取的所有有用信息，并以一种可以被模型处理的格式（通常是高维向量）表示。

这个嵌入然后被传递给解码器。**解码器的任务是根据从编码器接收到的嵌入来生成翻译后的文本（目标语言）**。解码器也使用自注意力机制，以及编码器-解码器注意力机制，来生成翻译的文本。



编码器Encoder-Only：以谷歌的BERT为代表，百度的文心属于该系列。

- 集中于理解输入的内容，并做针对特定任务的输出。
- encoder-only类型的LLMs更擅长对文本内容进行分析、分类，包括情感分析，命名实体识别。

解码器Decoder-Only：以OpenAI的GPT、谷歌的Bard、Meta的LLaMA为代表。

- Decoder主要是为了预测下一个输出的内容/token是什么，并把之前输出的内容/token作为上下文学习。
- decoder-only类型的LLMs更擅长各种Q&A。

编解码器Encoder-Decoder：以清华大学的GLM、Meta的BART、谷歌的T5为代表。

- 主要用于NLP，即能理解输入内容NLU，又能处理并生成内容NLG，尤其擅长处理输入和输出序列之间存在复杂映射关系的任务，以及捕捉两个序列中元素之间关系至关重要的任务。
- encoder-decoder类型的LLMs充分利用了上面2种类型的优势，更擅长输出强烈依赖输入的，比如翻译和文本总结。



大语言模型LLM

■ GLM (General Language Modeling) 模型：全国产化大模型底座（操作系统级别）

- 由清华大学 KEG 实验室和智谱公司共同训练，结合了 BERT 和 GPT 的优势，采用统一预训练模型。能够较好地处理上下文信息，并且能够适应不同的任务和数据集。
- 在中文处理领域具有全球领先地位，具有问答、多轮对话和代码生成功能的千亿中英双语模型。
- Stanford报告的世界主流大模型评测：中国唯一入选模型，准确性、恶意性与GPT-3持平，鲁棒性和校准误差在所有模型中表现最佳。

■ GPT (Generative Pre-trained Transformer) 模型

- 是一种生成式预训练转换器，由 OpenAI 于 2020 年推出，GPT 模型的优点在于能够生成自然流畅的文本，并且能够处理较长的文本序列。

GLM和GPT等大模型发展预测：

- 1. 模型规模的扩大：**随着计算能力的提升和数据量的增长，GLM 和 GPT 等大模型的规模将进一步扩大，以实现更高的性能和更广泛的应用。我们可以看到万亿级甚至亿亿级的语言模型。
- 2. 模型结构的改进：**GLM 和 GPT 模型将持续优化其结构，以提高模型的效率和准确性。这可能包括尝试更多元化的注意力机制、引入外部知识、使用更复杂的编码器和解码器结构等。此外，模型们也将更加注重模块化和可扩展性，以适应不同场景和任务的需求。
- 3. 跨模态应用：**随着人工智能技术的发展，GLM 和 GPT 模型将不仅仅局限于文本处理，还将逐渐拓展到跨模态应用，如图像、语音等。这将有助于实现更丰富、更智能的人机交互。
- 4. 下游任务的优化：**GLM 和 GPT 模型将更加关注下游任务的优化，以提供更加专业化和细粒化的应用。这可能涉及到针对特定领域和场景进行预训练和微调，以实现更好的效果。例如，在医疗领域，GLM 和 GPT 模型可以用于疾病预测、医疗问答等任务。
- 5. 模型伦理和隐私保护：**随着 GLM 和 GPT 模型的广泛应用，模型伦理和隐私保护将成为重要的研究方向。这包括防止数据泄露、消除偏见、确保模型公平性等方面。
- 6. 模型可解释性和可靠性：**为了提高 GLM 和 GPT 模型的可靠性和可信度，未来研究将致力于提高模型的可解释性，使人们能够理解和信任模型的决策过程。
- 7. 人工智能与人类的融合：**GLM 和 GPT 模型的发展将促进人工智能技术与人类的融合，例如通过脑机接口等技术实现人机共生。

国内外LLM(大语言模型)发展路径对比

■ 国内外大模型发展路径存在一定差异，主要表现在**技术、应用场景、生态环境和发展速度**等方面：

- **技术发展较为领先**，其研究机构和高校在大模型算法、模型架构、数据预处理等方面**已有较深入的研究**。

国内外大模型技术的差异

- 技术方面相对落后，但**发展速度较快**，在不断缩小与美国的差距。

- toC：更关注**消费者和服务业**应用，其在**医疗保健、金融科技、电子商务**等领域的应用较为突出。

应用场景的差异

- toB：更侧重于**政务、工业和企业应用**，其在**工业制造、金融、医疗**等领域应用较为广泛——toC、toB同时进行。

- AI生态环境较为开放，其数据开放程度较高，且**拥有较多的开源社区和AI平台**，使得其大模型技术得到更加广泛的应用和推广。

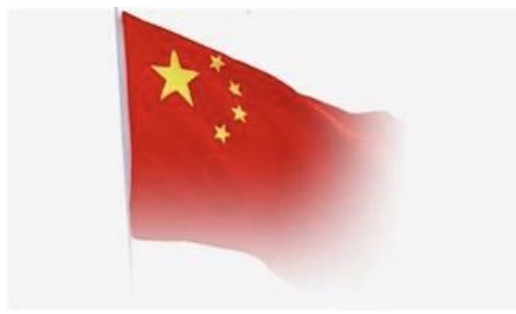
生态环境及开放性的差异

- AI生态环境相对较为封闭，对**数据安全和隐私保护的要求较为严格**，这也限制了大模型技术在一些领域的应用。

- 技术发展速度较快，其**研究和应用规模较大**，且拥有较多的创新型企业 and 研究机构。

发展速度及规模的差异

- 技术发展速度虽快，但其在**创新性和开放性**方面仍需进一步提高。



■ OpenAI发布会上最新亮点:

GPT-4 Turbo

更大的输入长度、更高的处理能力和灵活性、
更强的语义理解能力、更快的运行速度

GPTs

下调了API的价格；让人们能用**自然语言**构建定制化GPT，然后可上传到「GPT Store」中

GPT Store

以**GPT为中心的生态**，旨在集成和赋能AI生态伙伴，提供更好的人工智能解决方案

目前国内的**自定义大语言模型（GPTs）**主要有以下不足



自定义化自由度不够

不能使用一个基底模型衍生出多种自定义的助手，问题处理能力弱，行为化单一。



多模态融合能力不理想

非结构化数据的位置对应关系的研究仍有待提高



模型所适配的数据单一

- 知识库支持的类型太少，只能适配txt文件、pdf等文件；
- 需要用户自己对数据进行处理以后上传；
- 文件处理的大小有限；
- ✓ 应优化多种**结构化与非结构化数据**处理，如PPT、doc等



生态不够完善

- 没有将模型集成到插件里，不能和其他应用进行联动；
- ✓ 需把每一个定制化大模型以插件的形式，实现模型与接口间的联动。





科研

政务

医疗

传媒

社区

- 大语言模型强大的泛化能力可在大量数据中识别出关键模式及关系，发现新的研究趋势，可以帮助科研人员更好地理解 and 解释复杂的数据和现象，自动化预测及生成实验结果，提升工作效率及质量。

文献分析

可以对大量文献进行自然语言处理和情感分析，帮助研究人员更好地理解和分析人类语言行为，例如文本分类、情感分析、信息抽取等，更好地管理和组织大量的数据和文献。

机器翻译和自动写作

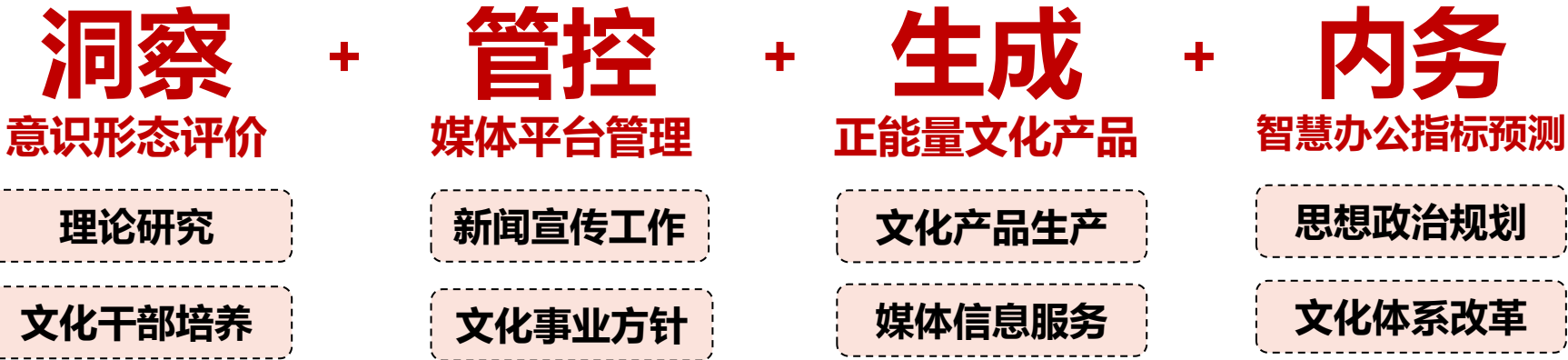
可以进行机器翻译和自动写作，帮助研究人员实现跨语言的科研合作和创新，例如对学术文献的自动翻译和自动写作。

舆情分析和社会热点监测

可以对社交媒体等网络数据进行分析 and 挖掘，帮助研究人员了解社会舆情变化及社会热点话题的演变，例如开展事件监测、社会网络分析等，发现新的研究趋势和领域。

化学及生物研究

如EPFL团队开发的基于大型语言模型的化学引擎 ChemCrow，可以使用 17 种工具，这些工具为其提供有关分子和反应属性等知识；在生物医学领域，可以用来分析基因组数据，以发现新的疾病治疗方法和药物。



驱动内容生成

- **公文起草**
根据设定模板及工作内容记录自动起草工作文件或绩效报告等。
- **设计制作**
根据文字内容生成图像，包括版面设计、插图、封面设计等。
- **文稿初加工**
对文稿进行初加工，并留痕。包括但不限于排版审核、错字修改、内容审核（涉黄涉政等）、文字校对；

提升处理非结构化数据的能力

- **图书分类**
通过内容识别功能提取图书关键词并生成摘要，辅助完成图书分类。
- **发文初审**
包含文字检查(错字标点等)，内容初审(政策事实引文等)，排版初审(人名顺序等)。
- **版权审查**
对文字、内容、代码等进行查重，完成著作权的审核；

提升人机交互能力

- **智能问答**
从传统的单向交流，变为对话式的双向交流。使得复杂工作，可以通过与智能助手的对话实现。
- **智能检索**
构建智能检索系统，集成至办公系统帮助用户快速查找和获取相关信息，提高信息检索效率。

- 大语言模型在医疗领域的应用仍处于**探索和研究**阶段，需要进一步的研究和实践来验证其可行性和效果；
- 由于医疗领域的复杂性和敏感性，使用大语言模型进行医疗应用时需**遵循相关法律法规和伦理规范**。

医学图像分析

可以处理医学图像数据，如X光片、CT扫描和MRI图像等。这有助于医生更准确地诊断疾病和评估病情，提高诊断的精准度和效率。



医学教育

通过对大量的医学文献和病例的学习，为医学生和医生提供教育和培训，有助于提高医学生的培养质量和医生的诊疗水平。



临床诊疗决策辅助支持

通过对医疗文献的学习，理解疾病的症状、诊断方法和治疗方案等；为医生提供参考和辅助，帮助医生做出更准确的诊断和治疗方案。



个性化治疗建议

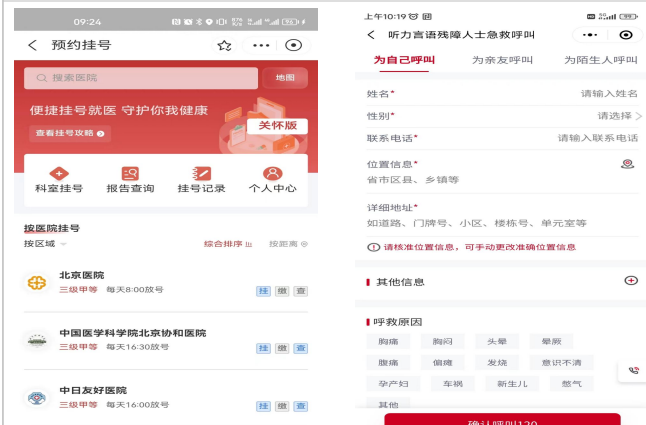
根据患者的病历、基因组和生活习惯等信息，为患者提供个性化的治疗建议，有助于提高治疗效果，减少副作用，提高患者的生活质量。



大语言模型未来发展方向探索：医疗（2/2）

- 将信创化大语言模型与现有医务数字化平台打通，真正实现由数字化向智能化转型，减轻就医压力。

京通健康



预约挂号老年版

听力言语残障服务

- “京通健康-北京114预约挂号平台以数字化手段打通医疗服务“最后一公里”，用实际行动解决群众看病就医急难愁盼问题；
- 提供**270家医院**的预约挂号服务，基本覆盖全市重点三级医院。
- 增加号源总供给量，**向特色专科倾斜**，并优化查询功能。

999非急救医疗服务平台



一键呼救，实时定位

医生上门检查

康复服务

康复设备租赁

居家服务

送药上门



- 不出家门，一站式服务
面向**非急救医疗转运、移动医院巡诊巡检、上门服务、送医送药上门**等非急救医疗服务场景，定制化设计运行环境，提供**云+网+平台+终端+安全**的一揽子服务。

5G救护车



院内远程会诊

电脑远程指导

手机远程会诊

软件平台



更多
息化平台

身份信息
人脸识别
考勤打卡

支付系统
定位系统
药品扫描

大语言模型未来发展方向探索：智慧社区3.0

智慧安防

智慧家庭

智慧生活

数字治理

数据服务

智慧养老

- 智慧养老终端设备
- 智慧养老监管服务平台
- 助残养老核心服务
- 辅助质量监管



面向社区/园区老龄群体、残障群体以及特需群体提供智能化、科技化关爱手段，由户外活动到室内行动再到与人交互，从常规情景到突发状况，全方位、全流程支持社区/园区“有爱”而“无碍”，让服务更有温度。

统一规划、统一标准、统一设计、统一建设、统一运维

将着眼于社区分类建设，统一满足政府建设要求，打造市级示范类社区建设。

为后续的进一步的新型智慧社区拓展计划留下充足的空间，随时可在现有的建设基础上进行持续的升级改造。

大语言模型未来发展方向探索：传媒（1/3）

智能打标 助力扫黄打非

- 通过对文本、图片、音/视频等信息进行内容理解，根据标签库完成打标，实现对内容风险的精细化辨别。
- 图像识别、语义理解、音频识别、标签库、策略配置

态势感知 辅助运营监测

- 通过监测到的事件发展态势进行预测，向决策者展示类似历史案件并提供处理建议。可以根据最新政策灵活调整处置策略。
- 数据采集、数据预处理、数据分析、态势预测

思想文化阵地管控平台

公众信息洞察大模型

运营监测指挥平台

智能监管 辅助游戏防沉迷

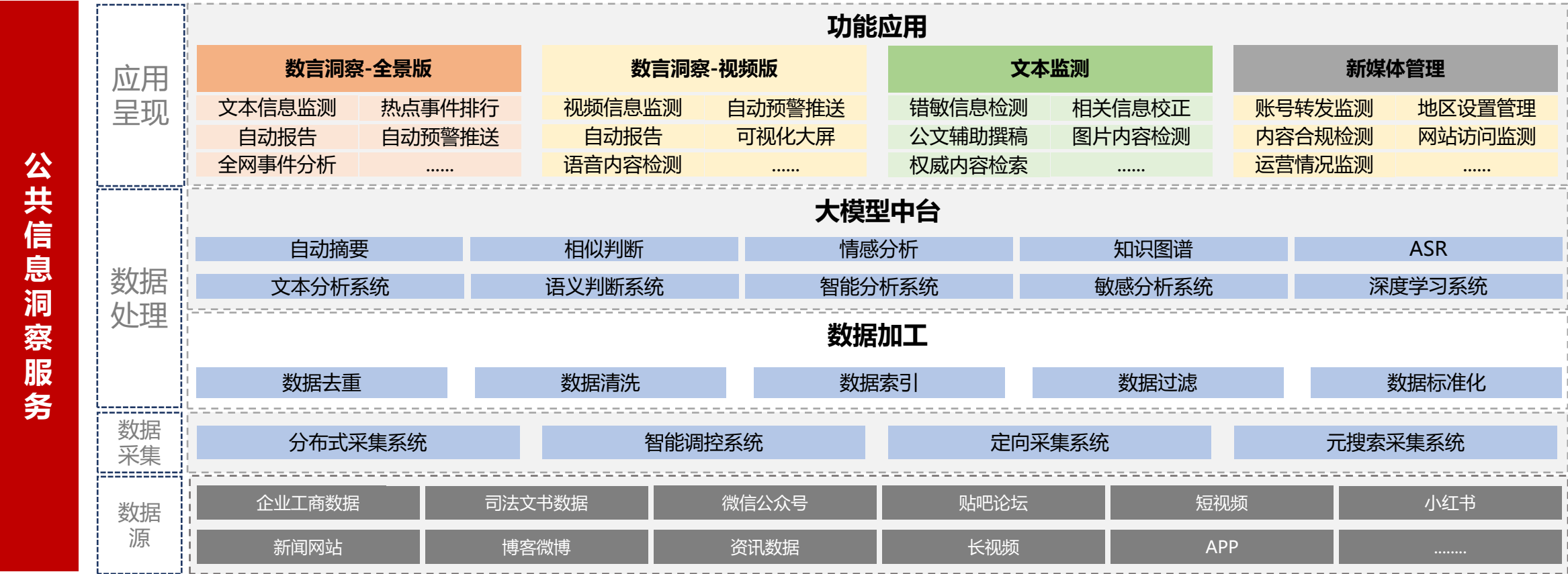
- 通过对玩家游戏时长、频率及消费等数据进行行为分析，实行监控，防止沉迷；
- 通过人脸识别、语音识别等技术预防未成年人借用成年人账号进行游戏。
- 防沉迷系统 家长监控

热点溯源 辅助运营监测

- 通过追踪、采集、分析热点事件相关数据及信息进行溯源，并将结果以时间线、人物关系图等可视化形式进行展现。
- 热点事件识别 溯源模型 可视化展示

信息洞察服务矩阵+智谱清言大模型

- **数言洞察**是基于对**互联网信息采集、文本挖掘和智能检索**，及时发现并快速收集所需的网络信息，并通过**大模型**自动采集、智能过滤、自动聚类、语义识别、统计分析等技术与应用，实现对**社会热点话题、突发事件、重大情报**的快速识别和**定向追踪**。对有较大影响的事件实现快速发现、快速处理，实现**网络动向的及时掌握并辅助决策**，高效进行**正面引导和宣传**。



- **AI有效内容生成平台**是一款基于**多模态大模型**的智能化内容创作工具，旨在帮助用户**快速生成高质量的评论、转发文案等**。用户只需输入关键词、上传短视频、或直接想，即可在短时间内获得与之相关的优质文案。该平台面向**正能量社交媒体推广**领域，具备智能化的创作方式和简洁的操作流程，**可充分提升公众的实时热点接受度和有效内容传播效率**。



实现过程

数据收集和预处理

- 收集大量与目标关键词相关的**短视频素材**；
- **使用NLP技术**对其内容抽取和分析，提取关键词、主题等信息；
- 进行分词、去除停用词、词性标注等预处理。

文案生成模型训练

- 基于**大规模的关键词和文本数据集**，训练一个文本生成模型，例如seq2seq模型、Transformer模型等；
- 基于**多模态大模型技术**，完成自动分析及预研。

文章生成流程

- 输入关键词，系统根据**关键词**选出合适的素材并进行预处理。
- 预处理后的文本传入训练好的大模型，**生成相应的文章文本**。

生成效果评估和优化

- 对生成的文案进行人工评估，**比对生成的文案与相关短视频的一致性、可读性、连贯性等指标**。
- 根据评估结果，优化模型的训练和生成过程。

用户界面设计和部署

- 设计一个用户友好的界面，用户可以**输入关键词**并查看生成的文案。
- 部署服务器和相关组件，确保系统的高**可用性和性能**。

技术优势

01快速生成

- 用户只需输入**关键词**或**上传短视频**，即可快速生成相应的优质评论、转发文案。

02智能化

- 基于海量数据，采用**先进的NLP技术和大模型训练基础**，自动分析关键词、短视频，避免繁琐人工创作。

03多样的文案风格

- 具备多种文案风格，包括**幽默、正式、煽情等**，以满足用户不同的创作需求。

04实时更新

- 实时更新数据库，收录最新的流行词汇和热门话题，确保生成的文案**更拟人化、更具吸引力**。

05易用性

- 界面简洁易用，操作便捷，无需专业技能即可使用，**降低了使用门槛**。

把握要点 健康发展

1

数据安全和隐私保护

大模型需要大量的数据进行训练，但这些数据往往包含着**用户的隐私信息**。因此，需要采取**有效的措施**来保护用户隐私和数据安全。

2

技术标准和互操作性

由于涉及多领域的技术，如计算机视觉、自然语言处理、语音识别等，因此需制定**统一的技术标准**和保证不同模型之间的互操作性，以便更好地**实现跨领域应用**。

3

伦理和社会责任

大模型的应用可能会对人类社会产生重大影响，如就业、隐私、安全等方面的问题。因此，需**遵循伦理规范和社会责任**，确保大模型的应用符合人类社会的利益。

4

自主可控和国产替代

信创领域的核心是**实现信息技术自主可控**，**规避外部技术制裁和风险**。因此，需时时关注**自主可控和国产替代**的问题，推动国内大模型技术的发展和**应用**。

一. 算力时代到来，聚焦信创化体系

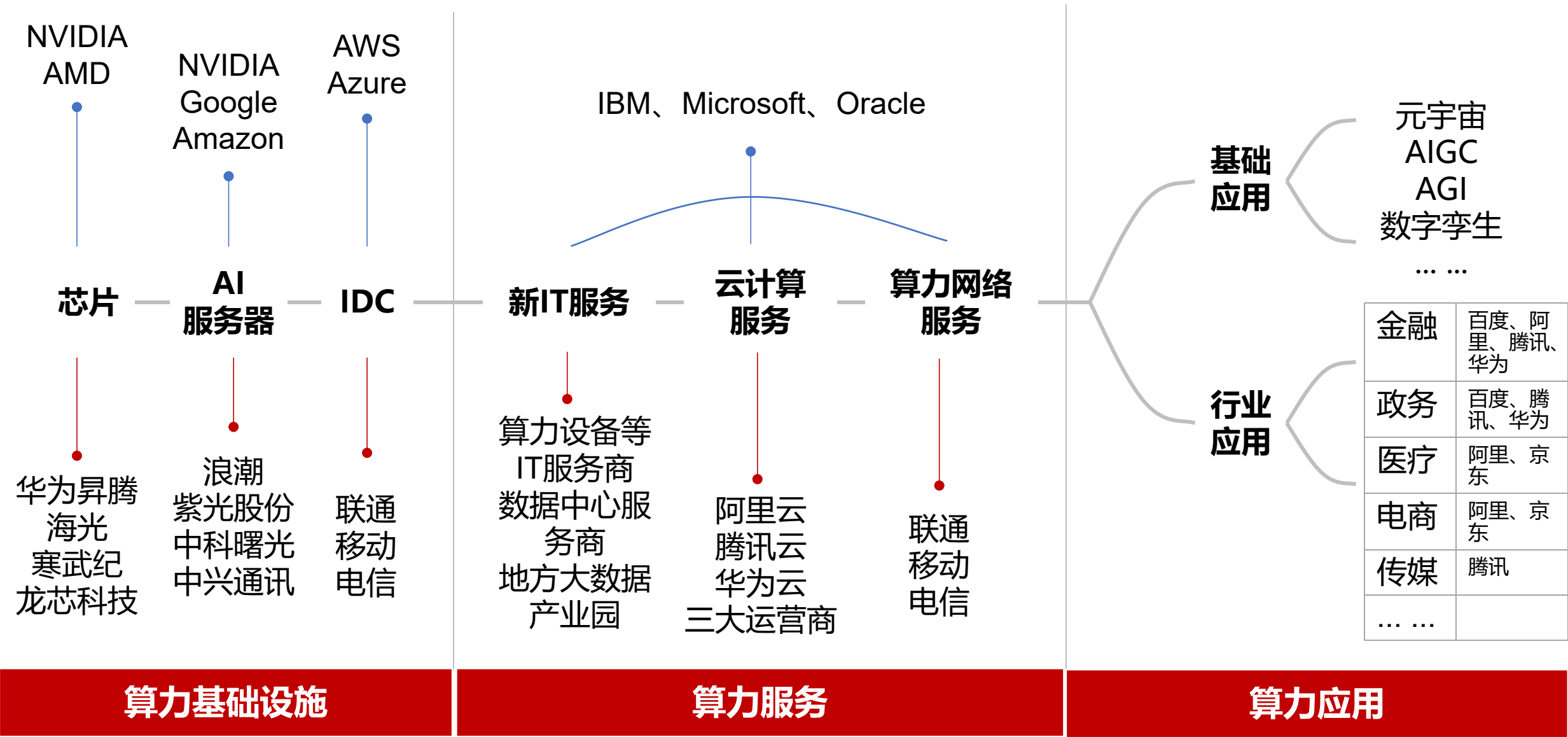
二. 国产硬件革新，对标英伟达算力

三. 模型软件齐放，适配信创化平台

四. 行业模型并进，toB & toC共生

五. 安全体系保障，信创化健康发展





信创之路：打造网络强国“算力+算法”信创优势

- 不同的开源模型采用了不同的架构和算法，因此在性能和适用场景上可能存在差异。对于特定任务和行业来说，找到**最适合的模型架构**至关重要。然而，在众多开源模型中找到“**最优解**”需要耗费大量时间和资源。
- 联通与**清华智谱、华为、百度**等公司在AI领域创新合作，在项目建设、标准制定和人才培养等方面助力我国AI芯片产业发展。

合作生态优势

- **优化资源配置**：与各行业领军企业、政府部门等开展深度合作，实现新兴技术与各领域的融合发展，拓宽产业布局，**整合各方资源配置资源**。
- **创新技术研发**：与科研机构、高校等合作伙伴共同开展前沿技术研究，如算力、大模型、大数据、人工智能等，推动技术突破和创新。**共同研发高性能、降低低能耗的算力产品**。
- **打造开放平台**：打造开放的合作平台，吸引国内外优秀企业、科研机构共同参与，**共建公共智算平台**。
- **促进产业应用**：推动算力在各个领域的应用落地，**助力产业发展和升级**。
- **建立行业标准**：探讨并制定相关技术标准和规范，**引导行业健康有序发展**。

通用大模型

清华智谱	昆仑天工
文心一言	华为盘古
通义千问	讯飞星火
智源悟道	...

行业大模型

盘古气象	ChatLaw
鹏程神农	Code GeeX
空天灵眸	阿里犀牛
360智脑	...

基础大模型（基座）

国产		国产		
GLM		GPT	T5	BERT
		LLaMA	BART	BARD



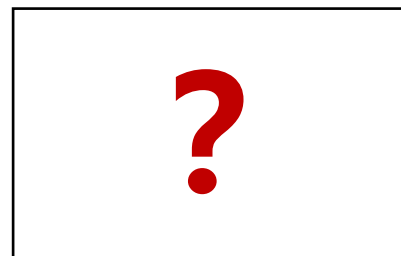
智能手机操作系统



支付平台



云服务平台



AI平台

AI平台化核心：
“大模型+微调” AI范式

■ AI平台既是一项**技术平台**，也可以**孵化更多新型应用**，可以有效打通底层算力和上层应用，某种程度上是**AI应用的基础设施**，这种新形态**对各个领域都将产生重大影响**，加速数字化对传统服务的取代。

- AI平台企业将大幅压缩现有碎片化 AI 厂商的市场空间。
- 围绕 AI 运维、低代码、数据服务等促成平台模型落地的工具组件将快速涌现。
- AI 平台企业将对既有内容市场、IT 服务市场产生高度影响。
- 全面智能化发展将再次将数据要素、数字资产提上日程，业界更加呼唤可信数据服务。
- AI 产业的变革可能会进一步促成算力服务市场的转变。

碳达峰碳中和

绿色新型数据中心

绿色技术产品

高效IT设备

高效制冷系统

高效供配电系统

高效辅助系统

新能源

分布式光伏发电

燃气分布式供能

清洁/可再生能源

绿色管理

绿色设计

绿色施工

绿色采购

绿色运营

芯片类型

最大功率

CPU

500W

GPU

(以DGX A100为例)

6,500W

2020年中国数据中心能耗总量为**939亿**

千瓦时，碳排放量为**6464万**吨；

预计到2030年，能耗总量将达到**3800亿**

千瓦时，碳排放将超过**1.9亿**吨（增长率超300%）。

数据安全

数据安全已被提升至国家立法层面，《国家安全法》《网络安全法》《数据安全法》及《个人信息保护法》等共同构筑了数据安全保护的法律法规框架体系。

推动算力发展过程中，数据流动和安全发展的矛盾将逐渐显现，必须严守数据安全防线，规避数据泄露、数据贩卖、数据滥用等风险。

- 打造数据安全治理平台，推进人工智能数据安全标准制定。

算法安全

构建互联网信息服务算法安全监管体系，《互联网信息服务算法推荐管理规定》，为互联网信息服务算法综合治理提供了重要法治保障。

在人工智能技术快速发展、大模型爆发的背景下，算法应用中存在的违法和不良信息传播、侵害用户权益、操纵社会舆论等乱象问题屡次出现。

- 建立健全算法安全治理机制，构建完善算法安全监管体制，推进算法自主创新。

算力安全

保障算力安全需加强重视，在国家政策方面，重视程度相对不足，多以地方政府级别政策为主（上海、成都等）。

面临国际高端技术封锁的困境，包括高端芯片、数据中心、算力网络等算力基础设施，以及技术路线生态的自主可靠与安全可靠亟需加强关注。

- 通过产学研联合攻关提升算力自主可控能力，新型数据中心建设也需超前布置，提高承载能力。

数字信息基础设施运营服务国家队
网络强国数字中国智慧社会建设主力军
数字技术融合创新排头兵