DeepSeek 开创行业新纪元

——开启行业变革，驱动技术革新

目录

DeepSeek发展历程

- DeepSeek成立于2023年7月17日，由**杭州深度求索人工智能基础技术研究有限公司**开发
- 2024年1月5日，其发布第一个同名AI大模型**DeepSeek LLM**
- 2024年12月26日，开源**DeepSeek V3模型，671B**参数量的MoE模型，激活参数量为**37B**
- 2025年1月20日，开源**DeepSeek R1**推理大模型，火爆出圈，其性能超过OpenAI o1模型，且完全开源，调用成本降低了**90-95%**

AI + 国产+免费+开源+强大

deepseek

FIRE-FLYER

幻方 | 人工智能

2019 **成立幻方人工智能基础研究有限公司**

2019 **萤火一号集群500卡后增加1100加速卡 2亿投资**

2021 **萤火二号集群 10000卡 A100 10亿投资**

2022 **美国管制高端GPU芯片出口 萤火二号扩容翻倍**

2023 **成立DeepSeek 杭州深度求索人工智能基础技术研究有限公司**

2024 **DeepSeek V1**

2024 **DeepSeek V2 每百万token 仅1元**

2024 **DeepSeek V3**

2025 **DeepSeek R1-zero**

2025 **Janus多模态 R1**

DeepSeek大模型介绍

- DeepSeek大模型通过强大的计算能力和低延迟推理，优化资源利用效率，助力企业快速处理大规模数据，推动AI应用性能的突破
- DeepSeek大模型面向用户和开发者，提供智能对话、文本生成、语义理解、计算推理、代码补全等应用场景，支持联网搜索和深度思考模式，可上传并解析各类文件和图片中的文字内容

01

高效推理引擎

- 基于并行计算架构，显著提升推理速度和吞吐量

02

低成本推理优化

- 通过精细化的资源调度和运算优化，降低计算成本并提高效能

03

MIT开源协议

- 开放源码，支持自由修改与二次开发，允许基于模型输出进行蒸馏与迁移学习

模型版本	参数量	模型说明	模型定位
DeepSeek V3	671B	•MoE语言模型，总参数量671B，激活参数量37B参数，基于 14.8T tokens 训练，实现高效的推理和经济高效的训练	•专为通用型大语言模型，专注于自然语言处理(NLP)、知识问答和内容生成等任务
DeepSeek R1	671B	•DeepSeek-R1是基于DeepSeek V3 采用推理（Reasoning）方式基于强化学习训练的复杂推理模型，该模型在很多任务场景效果超过 OpenAI o1，业界影响力巨大	•专为复杂推理任务设计，强化在数学、代码生成和逻辑推理领域的性能
DeepSeek R1 LLM	1.5B-70B	•基于 DeepSeek R1 轻量化版本，采用蒸馏技术，参数量减少 30%-50%，通过 R1 生成的样本数据微调小型稠密模型，将 R1 的推理能力转移到更小模型中，实现更低计算资源消耗和高效推理	•适用于边缘计算、移动设备和资源受限算力环境

DeepSeek功能与用途

DeepSeek

自然语言类	语义分析	文本分类	知识推理	···
客服问询类	智能客服	政务问答导办	差旅问答	···
岗位助手类	风险识别	会议纪要	会议预订	···
资料审核类	在线审核	合规审核	投标审核	···
文本生成类	文本创造	摘要与改写	结构化生成	···
常规绘图类	SVG矢量图	Mermaid图表	React图表	···
编程代码类	代码生成	代码调试	技术文档处理	···

DeepSeek & ChatGPT 比较表

项目	DeepSeek	ChatGPT
开发公司	深度求索（中国）	OpenAI（美国）
成立时间	2023年	2015年
开发时间	较新，技术仍在快速发展中	较成熟，已迭代多个版本(GPT-1 到 GPT-4)
开发成本	未公开，可能投入大量资源于 AGI 研究	数亿美元(包括硬体、数据和研发)
训练成本	558万美元(DeepSeek-V3)	10 亿美元(GPT-4o)
目标	实现通用人工智慧(AGI)	开发强大的自然语言生成模型
核心技术	深度学习、自然语言处理、多任务处理	GPT架构(Generative Pre-trained Transformer)
语言优势	中文处理优化	英文处理优化
开源情况	DeepSeek-R1(开源)、API(收费)	旧模型(GPT-2)开源，新模型(GPT-3、GPT-4)闭源
免费版	目前无明确免费版资讯	有免费版，但功能有限要排队
付费版	价钱尚未公开，可能按使用量收费	ChatGPT Plus：每月 20 美元(约 600 台币)
API价钱	0.14 美元(输入)	2.5 美元(输入)
应用场景	多任务处理、专业领域问答、中文环境	文字生成、对话系统、英文环境
对话能力	强调多轮对话和复杂问题解决	擅长生成连贯、自然的对话内容
文本生成能力	支持中文文本生成，质量高	英文文本生成能力极强
翻译能力	中文翻译优化	英文翻译优化
企业合作	可能专注于中国市场和企业合作	全球范围内合作，企业应用广泛
未来发展	专注于 AGI，目标是更通用的 AI	持续优化语言模型，扩展应用场景
硬体需求	未公开，可能需高效能计算资源	需要大量 GPU 和高效能计算资源
数据来源	未公开，可能包含大量中文数据	来自网路文本、书籍、文章等多种来源
用户评价	尚在发展中，用户评价较少	全球用户评价高，尤其英文用户

DeepSeek 为什么大火

表面原因：

比肩ChatGPT O1的模型能力表现
+
极低的训练成本
+
开源普惠，免费商用

> **技术价格双重普惠**
> **技术创新行业进步**

先进性能	**物美价廉**	**开源开放**	**模型出色**	**四大创新**
DS-R1性能比肩ChatGPT O1模型	低算力高质量 每Token费用为O1的**1/10**	V3/R1模型开源免费商用	671B及1.5-70B多种选择	数据、模型、算力、硬件

深度影响：

打破算力至上的传统认知，推进工程优化

AI技术回到算法创新阶段，加速产业进程

重构AI生态加速推理落地，应用需求爆发

> **AGI与技术普惠**
> **促进生态多元化**

DeepSeek 相关理解误区

绕过了CUDA壁垒

- **事实**：基于NVIDIA 的PTX 层优化 (H800 带宽补救方案)
- **真相**：PTX 属于CUDA技术体系，依赖 NV 驱动工具链

全面赶超国外

- **事实**：部分功能达国际主流水平 (如R1推理能力)
- **真相**：综合差距缩短至0.5-1年(原3年),芯片/生态/IT基础设置仍存在差距)

训练成本极低

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

- **来源**：V3 基础模型算力成本(278.8万 H800 小时×$2)
- **实情**：未含RLHF调优/人员/ 基建，第三方估算5-13亿美金

完全开源

- **现状**：开放模型权重，免费商用，引发国产化适配潮
- **局限**：训练框架/数据工程等核心信息未公开，与Llama系列模型开源形式一致。

目录

各国对DeepSeek的反应：限制、调查与拥抱

DeepSeek 爆火引发了全球关注，各国基于自身利益和立场采取不同措施

1 持谨慎态度，开展调查评估

- 英国、法国、德国、爱尔兰等国的相关监管机构以用户数据安全为由发起对 DeepSeek 的调查

2 积极拥抱

- 印度信息技术部长称赞 DeepSeek 是重大突破，政府计划将 DeepSeek 模型托管在本地 AI 计算设施上，希望通过学习 DeepSeek 模式推动本国 AI 技术的进步

3 采取限制和禁止措施

- 美国国会、海军、NASA 等多方面已禁止使用 DeepSeek，国会还提出法案，拟规定下载或使用 DeepSeek 为犯罪，最高可判 20 年监禁
- 意大利隐私监管机构要求提供用户使用数据问题的解释，后将 DeepSeek 从该国应用商店下架
- 加拿大、澳大利亚政府下令禁止在官方设备上使用 DeepSeek 产品服务
- 韩国多部门禁用且要求从 2 月 15 日起暂停新用户下载

国内云厂商效应

华为云昇腾云服务

- 2月1日，华为云团队，经过连续奋战，成功推出了基于华为云昇腾云服务的DeepSeek R1/V3推理服务，这一消息无疑为国产AI的发展注入了强劲动力

腾讯云HAI平台

- 2月2日，腾讯云宣布DeepSeek-R1大模型已一键部署至其“HAI”平台上。开发者仅需3分钟就能接入调用这一先进的大模型，极大地加速AI应用的开发和部署进程。

阿里云PAI Model Gallery

- 2月3日，阿里云微信公众号宣布，阿里云PAI Model Gallery支持云上一键部署DeepSeek-V3、DeepSeek-R1。在该平台上，用户可以简化模型开发流程，为用户带来更高效便捷的AI开发和应用体验

百度智能云

- 2月3日，百度智能云宣布，DeepSeek-R1和DeepSeek-V3模型已在百度智能云千帆平台上架，同步推出超低价格方案，并提供限时2周的免费服务。

字节云

- 2月4日，字节跳动的火山引擎宣布全面支持DeepSeek系列模型，支持V3/R1等不同尺寸的DeepSeek开源模型

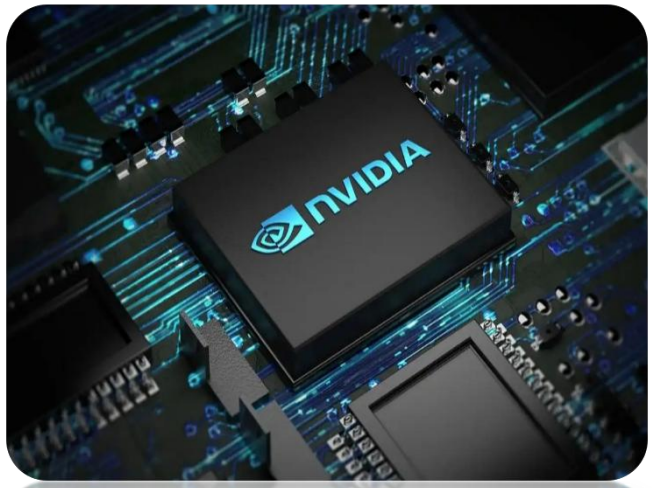
京东云

- 2月4日，京东云已正式上线 DeepSeek-R1和 DeepSeek-V3模型，支持公有云在线部署、专混私有化两种模式，供用户按需部署，快速调用

国际云厂商效应

2024年1月30日，微软宣布DeepSeek- R1模型可通过Azure AI Foundry和GitHub获取，并将在Copilot+ PC上运行，为开发者提供强大的AI支持。微软的合作使得DeepSeek能够快速进入国际市场，为全球开发者提供服务，推动了AI技术的全球普及

微软Azure AI Foundry



2024年1月31日，英伟达宣布DeepSeek- R1模型可通过NVIDIA NIM微服务使用，进一步拓展了DeepSeek的应用范围。英伟达的合作为DeepSeek提供了强大的硬件支持，提升了模型的推理性能，推动了AI技术在更多领域的应用

英伟达NIM微服务

2025年2月1日，亚马逊云科技宣布DeepSeek- R1模型已全面上线Amazon Bedrock和SageMaker AI平台，为开发者提供了更多选择。亚马逊云的合作使得DeepSeek能够更好地服务于全球开发者，推动了AI技术的广泛应用和发展

亚马逊云科技Bedrock

国际云厂商效应

谷歌（Google）高层公开评价

- CEO桑达尔·皮查伊多次称赞，认为DeepSeek团队“做得非常非常出色”，特别指出其开源模型的高效性和全球化价值，并强调这种创新强化了AI普及的潜力。他提到DeepSeek的成功促使谷歌意识到“为全球用户服务的模型必须非常高效”
- DeepMind负责人德米斯·哈萨比斯，称DeepSeek-R1模型是“所见过的来自中国的最佳AI作品”，肯定其工程能力对地缘竞争格局的影响，但也指出其技术仍基于“已知的AI技术”

设备厂商效应

浪潮信息

- 2月11日消息，浪潮信息宣布元脑企智EPAI企业大模型开发平台已全面接入支持DeepSeek大模型。通过元脑企智EPAI，用户能够将业务数据与DeepSeek大模型结合，深度开发模型潜力，快速实现本地化部署DeepSeek，构建准确率高、安全稳定的专属智能应用

中科曙光

- 2月3日，中科曙光宣布完成DeepSeek V3和R1模型与海光DCU的国产化适配并正式上线，推动了国产AI技术的自主可控发展

中科曙光的合作为DeepSeek提供了强大的国产硬件支持，提升了模型的性能和效率，推动了国产AI技术的快速发展

新华三

- 2月21日，紫光股份在互动平台表示，紫光股份旗下新华三正式发布灵犀大模型一体机（DeepSeek版），该产品包含纯享版和使能版两大系列，共计六大款型12款产品，覆盖14B至671B规模的DeepSeek大模型

运营商效应

中国电信

- 2月1日，中国电信天翼云率先宣布完成与DeepSeek的全面接入，成为国内首家实现DeepSeek大模型全栈国产化推理服务落地的运营商级云平台。天翼云自主研发的“息壤”智算平台与DeepSeek-R1/V3系列大模型实现深度适配优化，推出“天翼云+DeepSeek全场景解决方案，用户可在天翼云智算产品体系 —— 息壤-科研助手、天翼 AI 云电脑、魔乐社区、 “息壤”智算平台、GPU 云主机 / 裸金属开启智能新体验

中国移动

- 2 月 5 日消息，中国移动“移动云”宣布全面上线 DeepSeek，实现全版本覆盖、全尺寸适配、全功能畅用。在支持在线体验推理、API 调用、专属资源分钟级一键部署等能力基础上，将 DeepSeek 集成至移动云智能体平台

中国联通

- 2 月 5 日，中国联通公众号发文称联通云上架 DeepSeek - R1 系列模型，支持私有化和公有化场景

中国广电

- 2月21日，中广电移动公司宣布，已于2月17日上线广电5G DeepSeek智能客服系统，启动友好用户测试

DeepSeek对行业影响

DeepSeek引发了全球范围的关注，推动了AI行业从“算法驱动”模式转向新阶段，为各行业AI应用的落地提供了新的机遇

- 自2024年12月发布671B参数的V3和R1模型以来，DeepSeek的表现与GPT-4和OpenAI的最新版本相媲美。V3模型在2048张H800卡上经过56天训练完成，效率是同类MOE模型的**1.5至2倍**，训练成本比Llama模型下降了**11倍**

技术突破

从创新到突破：DeepSeek大模型未来之路

- **推动技术创新与应用**
  - DeepSeek通过MoE架构与FP8低精度训练，显著降低训练与推理成本，推动大模型行业向低成本、高效率发展
- **降低训练与推理成本**
  - DeepSeek-V3训练成本仅558万美元，大幅低于GPT-4的10亿美元，降低了企业和开发者参与AI开发的门槛
- **提升行业效率与竞争力**
  - DeepSeek低成本策略引发“价格战”，迫使大厂降价，推动大模型服务价格下降，促进技术普及与应用

挑战与机遇并行

引领AI行业变革的引擎，开创新时代的AI前沿

- 算力资源瓶颈
- 数据隐私与安全问题
- 算法优化与调整难度
- 市场竞争加剧，技术同质化
- 多模态数据处理挑战
- 人才短缺

开源赋能算力转型

开源生态与算力需求变革的双轮驱动

- **开源推动普惠化**
  - DeepSeek坚持开源，提供免费API和文档，打破技术壁垒，降低中小企业和个人开发者参与门槛，推动全球AI技术应用
- **算力需求结构性变化**
  - DeepSeek低算力依赖削弱高端GPU主导地位，算力需求从训练侧向推理侧倾斜，推动国内算力市场向绿色高效转型，优化数据中心资源配置

市场格局重塑

触发价格战

“低价策略引发市场震荡，重新定义行业定价”

- **低价策略：**DeepSeek低价策略引发行业“价格战”，迫使其他大厂跟进降价

  技术普及：降低大模型服务价格，推动技术普及，加速行业向高效、低成本方向发展

  资源优化：促使企业重新评估算力采购策略，优化资源配置，提升市场竞争力

推动行业竞争

“创新技术推动竞争加剧，激发行业活力”

- **引入新玩家：**开源模式引入更多二线公司与玩家，打破头部大厂垄断局面

  激发创新：激发小型AI企业创新活力，与大型企业展开竞争，推动行业多元化发展

  合作探索：促使企业加强技术研发与合作，探索AI发展新路径，提升行业整体水平

改变市场预期

“技术突破与成本优化，改变行业未来走向”

- **发展反思：**让市场重新审视AI大模型发展逻辑，不再单纯依赖算力与投资堆砌

  创新导向：引导企业关注软件架构与算法创新，寻求更高效、可持续发展道路

  资源调整：调整市场对算力硬件需求预期，促使数据中心优化资源配置，提升资源利用效率

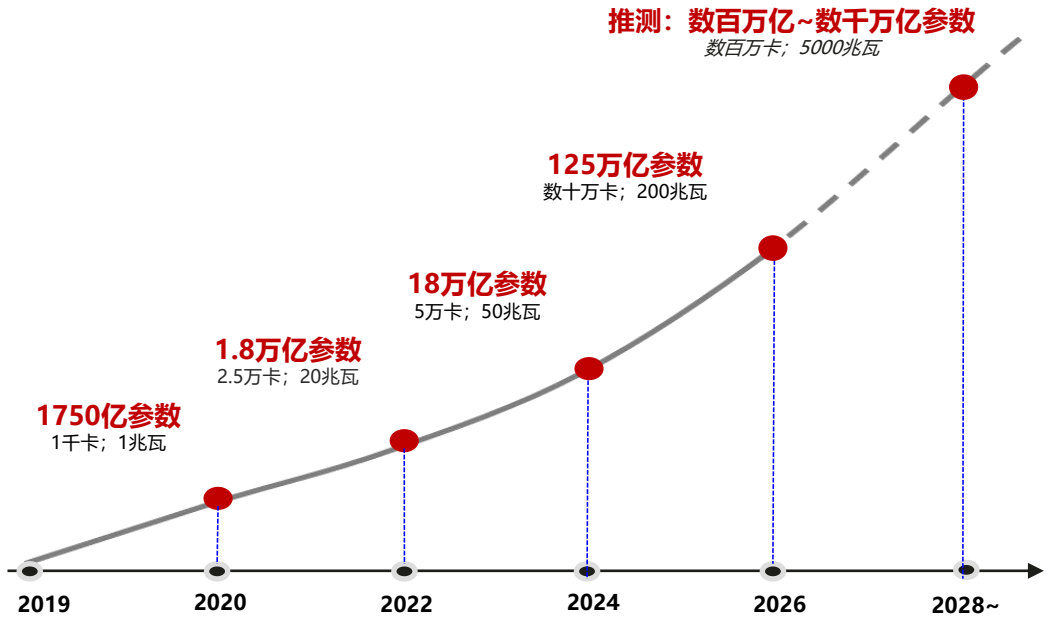
算力需求变革

算力需求结构调整

- **低算力依赖**：DeepSeek低算力依赖特性削弱高端GPU不可替代性，算力需求从训练侧向推理侧倾斜

  绿色高效：推动国内算力市场向绿色高效转型，促进数据中心优化资源配置，提升能源利用效率

  多元发展：长期来看，AI算力仍必不可少，但发展方向将更加多元化，依赖软硬件协同创新



对国产芯片影响

- **适配推动**：工信部要求模型厂商与国产芯片适配，DeepSeek开源模型助力国产芯片出货

  训练应用：推动国产芯片在训练领域应用，云厂商可通过规模效应降低成本，提升市场竞争力

  产业发展：为国产芯片厂商带来发展新机遇，打破国外巨头垄断，推动国产芯片产业发展

运营商新机遇

- **需求爆发**：集中式智算算力需求建设放缓，推理需求增加，分布式机房部署需求爆发

  资源利用：运营商利用现有资源与优势，提供算力服务与网络支持，拓展业务领域与市场空间

  合作创新：促进运营商与AI用户深度合作，共同探索AI应用与服务新模式，提升用户价值

下游应用与合作的拓展

01

广泛的应用场景

- DeepSeek模型在教育、医疗、金融、政务等多个领域广泛应用，开源模型对不擅长做基础大模型，但落地场景清晰的公司是利好，例如：
- 在医疗领域，DeepSeek协助医生进行疾病诊断，可提高诊断的准确性和效率，为医疗行业智能化转型提供了有力支持
- 在金融服务领域，它能帮助金融机构进行风险评估和智能投顾，推动了金融服务的智能化升级

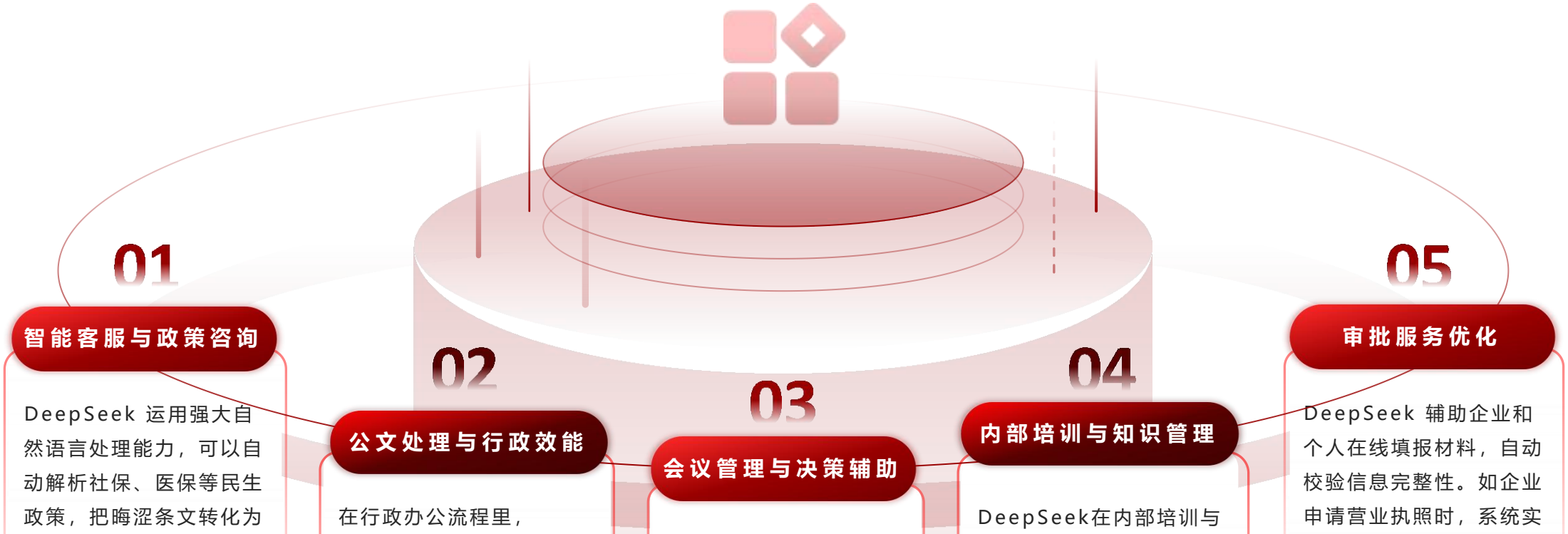
02

下游应用合作展望

- 下游应用合作展望方面，DeepSeek将持续推动AI技术的开源化、普及化和多样化发展。随着技术的不断进步，DeepSeek将在更多领域发挥重要作用，为各行业的智能化转型提供强大支持
- DeepSeek将需要继续优化其模型性能，降低训练和推理成本，推动AI技术在更多行业的应用落地，促进各行业的智能化发展

目录

DeepSeek行业通用业务场景



01

智能客服与政策咨询

DeepSeek 运用强大自然语言处理能力，可以自动解析社保、医保等民生政策，把晦涩条文转化为易懂的文本或语音回复，助力政策精准解读与咨询

02

公文处理与行政效能

在行政办公流程里，DeepSeek 借助强大分析能力，支持政策解读、文件智能核对，帮工作人员快速准确完成信息核对与数据统计，减少失误，提升公文处理效率与办公质量

03

会议管理与决策辅助

DeepSeek 革新会议管理与决策流程，会后分析录音或文本生成结构化摘要，提炼关键信息，提取待办事项并明确责任人与时间节点，提升执行效率

04

内部培训与知识管理

DeepSeek在内部培训与知识管理方面可搭建涵盖业务知识库、政策法规、业务流程、案例分析等内容，为新员工提供实时问答服务，助力其快速掌握业务流程

05

审批服务优化

DeepSeek 辅助企业和个人在线填报材料，自动校验信息完整性。如企业申请营业执照时，系统实时提醒遗漏必填项，减少重复提交麻烦，减轻人工审核压力

DeepSeek省内私有化部署业务场景(1/3)---经济版

使用场景

➢ **基础文本处理与生成**

内容编写：撰写演讲大纲、文章、报告等
文本摘要：生成新闻、报告或长篇文章的摘要
语言翻译：将一种语言的文本翻译成另一种语言
情感分析：分析文本中的情感倾向，如正面、负面或中性
命名实体识别：从文本中提取人名、地名、组织名等实体

➢ **数据分析与处理**

总结工作报告：生成周、月、季、年工作报告
数据分析报告：根据数据生成分析报告或摘要
财务数据分析：读取财务数据并生成分析报告
税务合规分析：根据法规知识和数据生成税务合规报告。

➢ **智能对话与交互**

智能对话：与用户进行多轮对话，理解意图并给出回答
客服应答：提供7×24小时自动化客服支持。
情感智能分析：在对话中分析用户情绪并做出相应回应

产品配置

预置模型：

DeepSeek R1/V3-671B(FP8)

DeepSeek-R1-Distill-Llama-70B

DeepSeek-R1-Distill-Qwen-32B

预置引擎：sglang、vllm、ollama

CPU： Intel 8352V *2

内存： 64G DDR4*16

系统盘： 480GB SATA SSD *2

数据盘： 7.68T U.2 NVME SSD *2

网卡： 双口10G光网卡*1

GPU：RTX4090 24G显存 *8

阵列卡： RAID卡

电源： 冗余电源

产品功能

公共服务力	
UI助手	可选择预置：OpenWebUI、ChatBot实现大模型的统一门户，提供AI原生用户体验，支持PC和APP端，支持语音交互
知识库创建	上传知识库、文本切分、文本块向量化、知识库管理的分级分权
知识库推理	多模型选择、对话框支持的输入长度可配置、语音转文字、知识库支持多轮会话、知识库检索
推理会话	支持大模型会话，可保留会话历史，支持简单参数配置，预置提示词（指令）推荐，支持挂载知识库进行问答
智慧应用	智能对话、知识问答、智能文档、文档阅读
智能应用	智能PPT生成、多语种翻译
专家服务	AIGC应用定制开发服务、知识库构建专家服务

DeepSeek省内私有化部署业务场景(2/3)---标准版

使用场景

- **经济版所有使用场景**
- **专业领域应用**

数学问题解决，逻辑推理。
科研分析，处理复杂的科研数据。
金融风控与数据分析。
合同分析，提取关键条款。
循证医学查房、病历摘要总结、临床诊疗辅助决策。

- **复杂任务与多模态**

多模态任务预处理。
复杂推理任务，处理高精度、高复杂度的推理。
大规模语言建模。
多tokens预测训练目标。
多阶段训练方式，包括基础模型训练、强化学习训练和微调

- **高级应用**

复杂业务决策支持。大规模云端推理。
处理海量数据。

产品配置

预置模型：

DeepSeek R1/V3-671B(FP8)

DeepSeek-R1-Distill-Llama-70B

DeepSeek-R1-Distill-Qwen-32B

预置引擎：sglang、vllm、ollama

CPU： Intel 6530*2

内存： 64G DDR5*16

系统盘： 480GB SATA SSD *2

数据盘： 7.68T U.2 NVME SSD *2

网卡1： 双口10G光网卡*1

网卡2： 400G 单口 IB卡 *4

GPU: HGX-H20-8GPU 96GB显存 *1

电源： 冗余电源

产品功能

公共服务力	
UI助手	可选择预置：OpenWebUI、ChatBot实现大模型的统一门户，提供AI原生用户体验，支持PC和APP端，支持语音交互
知识库创建	上传知识库、文本切分、文本块向量化、知识库管理的分级分权
知识库推理	多模型选择、对话框支持的输入长度可配置、语音转文字、知识库支持多轮会话、知识库检索
推理会话	支持大模型会话，可保留会话历史，支持简单参数配置，预置提示词（指令）推荐，支持挂载知识库进行问答
智慧应用	智能对话、知识问答、智能文档、文档阅读
智能应用	智能PPT生成、多语种翻译
专家服务	AIGC应用定制开发服务、知识库构建专家服务

DeepSeek省内私有化部署业务场景(3/3)---旗舰版

使用场景

- **经济版使用全场景**
- **标准版使用全场景**
- **高级应用**

复杂业务决策支持。
大规模云端推理。
处理海量数据，数据挖掘。
AI绘画，生成高质量图像。
AI写作，支持多语言、多模态内容生成

- **私有化部署**

私有化API服务，支持企业级私有化部署。
本地部署，支持单卡或多卡GPU运行。
优化推理性能，通过GPU/CPU异构协同提升效率。

产品配置

预置模型：

DeepSeek R1/V3-671B(FP8)

预置引擎：sglang、vllm、ollama

CPU： Intel 8468V *2

内存： 64G DDR5 *32

系统盘： 480GB SATA SSD *2

数据盘： 7.68T U.2 NVME SSD *2

网卡1： 双口10G光网卡*1

网卡2： 单口400G IB卡*4

GPU： **H100 SXM5 8GPU 80GB显存 *1**

电源： 冗余电源

产品功能

公共服务力	
UI助手	可选择预置：OpenWebUI、ChatBot实现大模型的统一门户，提供AI原生用户体验，支持PC和APP端，支持语音交互
知识库创建	上传知识库、文本切分、文本块向量化、知识库管理的分级分权
知识库推理	多模型选择、对话框支持的输入长度可配置、语音转文字、知识库支持多轮会话、知识库检索
推理会话	支持大模型会话，可保留会话历史，支持简单参数配置，预置提示词（指令）推荐，支持挂载知识库进行问答
智慧应用	智能对话、知识问答、智能文档、文档阅读
智能应用	智能PPT生成、多语种翻译
专家服务	AIGC应用定制开发服务、知识库构建专家服务

DeepSeek天翼云私有化部署场景(1/4)---产品类型

- **产品体系：**针对大模型微调训练、推理场景，提供一站式交付、开箱即用的软硬集成的一体机，分为推理一体机、训推一体机三大类产品
  - **推理一体机（DeepSeek版）**：基于DeepSeekR1/V3模型（32B、70B、671B等不同尺寸模型），结合息壤智算推理平台，为客户提供开箱即用的DeepSeek模型推理服务。
  - **训推一体机（DeepSeek版）：**基于DeepSeekR1/V3模型（7B、32B等不同尺寸模型），结合息壤智算训练平台，为客户提供全参微调和Lora微调等一站式服务

DeepSeek天翼云私有化部署(2/4)
---推理场景

- 服务期价格 = **一体机规格标准资费*数量** + **机柜标准资费** + **一体机规格维保标资费*数量**；第四年起每年将收取前三年总费用的 **10%**作为续期费用
- 后续进行扩容时，同样仅对扩容的服务器节点数量进行独立收费。（训练、模型参数调优、网络、机架、运维、培训）

系列		DeepSeek推理一体机-800TA2		
规格型号		专业版Pro-单机（推理版）	旗舰版Ultra-双机（推理版）	旗舰版Ultra+-多机（推理版）
DeepSeeK模型	适用模型规模	DeepSeek-R1-Distill-Qwen-32B 或者DeepSeek-R1-Distill-Llama-70B	DeepSeek-R1 671B（INT8）	DeepSeek-R1 671B（INT8）
性能	性能数据	32K上下文，32B:128并发：15 token/s 32K上下文，70B:64并发：14 token/s	**32K上下文, 671B, 64并发, 10 token/s** **（非高可用）**	**32K上下文, 671B,128并发, 10 token/s** **（高可用）**
NPU节点配置	CPU：4*鲲鹏920(48核2.6GHz)	1	2	4
	NPU：8*昇腾910B3			
	内存：24*64GB			
	系统盘：2*480G SSD			
	数据盘：2*3.2T NVMe			
管理节点配置	CPU：2*鲲鹏920(64核2.6GHz)/Intel同等性能	无	无	3
	内存：512GB			
	系统盘：960G SSD			
	数据盘：10T HDD			
交换机配置	GE交换机+25G交换机	1	1	2
	400G交换机	无	1	1
软件		息壤一站式智算服务平台-推理平台		
预售参考价	一次性付费（3年含维保）	284万	669万	根据实际部署情况定制
	分期付费（3年含维保）	100万	243万	根据实际部署情况定制

以上报价中不包含网络带宽费用

DeepSeek天翼云私有化部署(3/4)---训推场景

- 服务期价格 = 一体机规格标准资费*数量 + 机柜标准资费 + 一体机规格维保标资费*数量；第四年起每年将收取前三年总费用的 10%作为续期费用
- 后续进行扩容时，同样仅对扩容的服务器节点数量进行独立收费。（训练、模型参数调优、网络、机架、运维、培训）

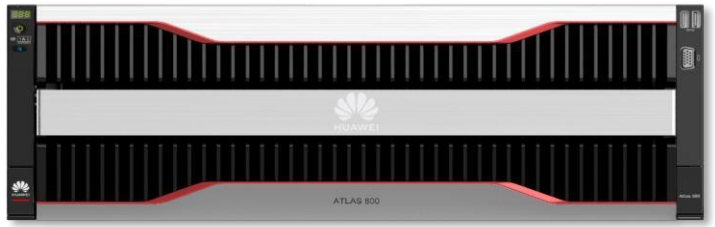
系列		DeepSeek训推一体机-800TA2		
规格型号		专业版-单机（训推版）	专业版Pro-双机（训推版）	旗舰版Ultra-多机（训推版）
DeepSeeK	适用模型规模	DeepSeek-R1-Distill-Qwen-7B及以下模型微调	DeepSeek-R1-Distill-Llama-14B及以下模型微调	DeepSeek-R1-Distill-Qwen-32B及以下模型微调
AI算力@BF16	推理性能数据	2.5PFLOPS	5PFLOPS	10PFLOPS
NPU节点配置	CPU：4*鲲鹏920(48核2.6GHz)	1	2	4
	NPU：8*昇腾910B3			
	内存：24*64GB			
	系统盘：2*480G SSD			
	数据盘：2*3.2T NVMe			
管理节点配置	CPU：2*鲲鹏920(64核2.6GHz)/Intel同等性能	无	2	3
	内存：512GB			
	系统盘：960G SSD			
	数据盘：30T SSD			
交换机配置	GE交换机+25G交换机	1	2	2
	400G交换机	无	1	1
软件		息壤一站式智算服务平台-训推平台		
预售参考价	一次性付费（3年含维保）	315万	730万	1413万
	分期付费（3年含维保）	113万	257万	500万

备注：结算价按照标准资费45折与天翼云公司进行结算；以上报价中不包含网络带宽费用

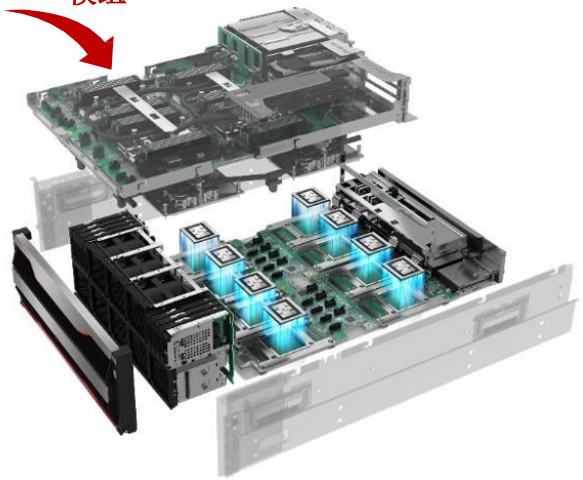
DeepSeek天翼云私有化部署场景(4/4)---服务器配置介绍

天翼云 China Telecom Cloud

Atlas 800T A2 服务器

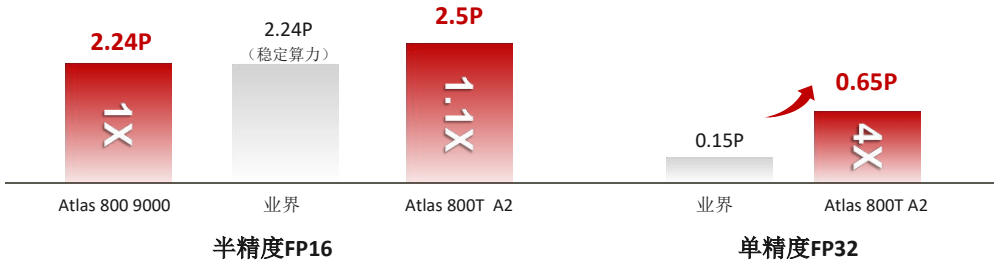


8 *  模组



关键特性	规格描述	
形态	4U高度训练服务器	
CPU	4 * 鲲鹏920	
NPU	8 * 昇腾910（64G）	
内存	容量：512G HBM；支持32个DDR4内存插槽 支持32个DDR4内存插槽	
内部互联	8NPU HCCS 全互联，互联带宽 392GB/s	
网络接口	NPU直出8 * 200G RoCE	
AI算力	半精度（FP16） **2.5** PFLOPS	单精度（FP32） **0.65**PFLOPS

对标业界，FP16及FP32算力全面领先



目录

数字政府领域

政务服务智能化

DeepSeek助力构建智能客服系统和智能审批应用，实现政策咨询、业务办理自动化与智能化，提升服务效率。

多地12345热线接入DeepSeek，实现智能问答、填单、政策解读等功能，增强政务服务便捷性与高效性。

政府决策科学化

深度挖掘分析大数据，帮助政府发现社会问题、经济趋势，为政策制定提供科学依据

对政府决策进行模拟预测，评估政策方案效果，推动城市治理精细化、智能化

政务治理精准化

实时监测分析城市运行数据，发现薄弱环节和风险点，为政府提供预警信息

智能优化调度城市运行，提高城市管理效率和水平，实现精准治理

教育行业应用

教学内容和课程设置

教学方法与教学模式

教育技术应用

DeepSeek助力教育行业智能化教学与精准个性化学习

➢ 个性化教学

- 根据学生/学员学习进度、知识掌握程度等，提供定制化学习路径与内容推荐，实现因材施教

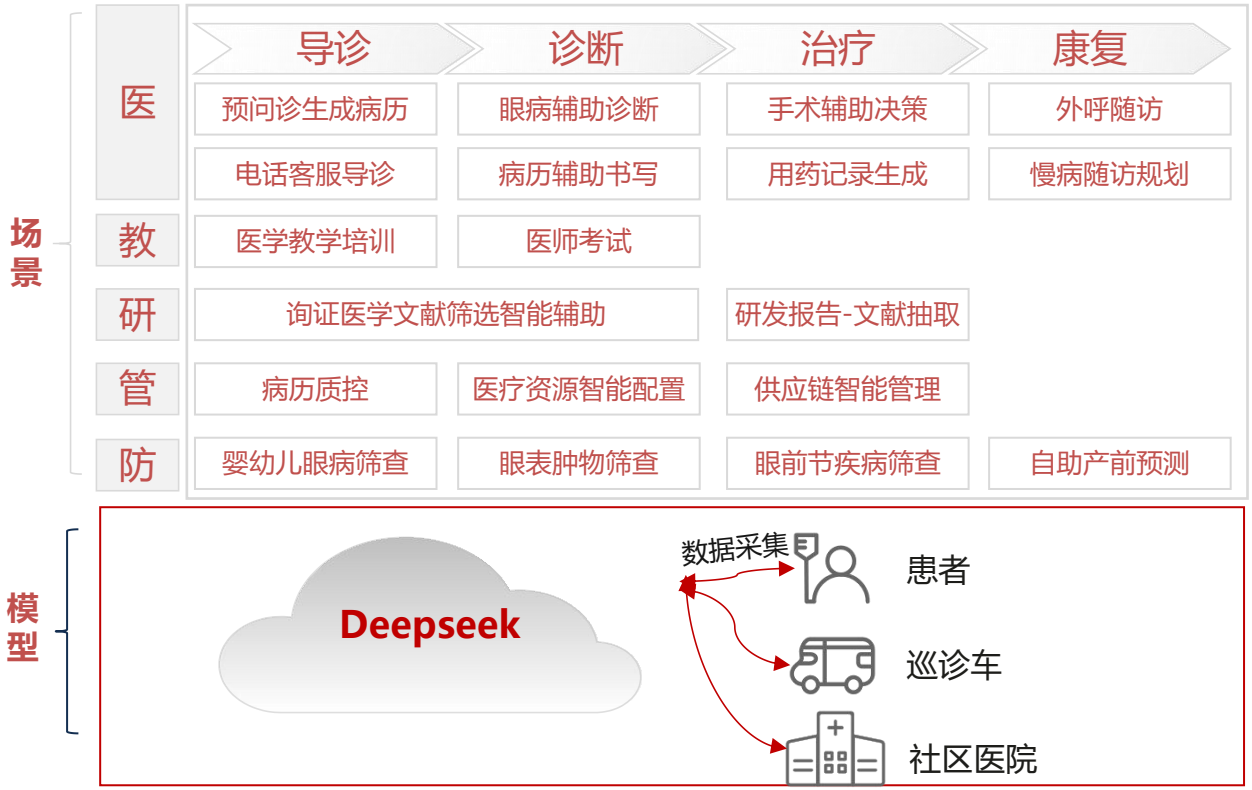
  协助老师快速生成教学课件、试题，丰富教育资源，提升教学效果

➢ 智能辅导服务

- 随时解答学生疑问，提供24小时不间断学习支持，为学生提供优质教育资源和服务

  利用AI技术，提升教育平台用户体验，推动教育行业智能化发展

医疗行业应用



疾病诊断辅助

- 学习海量医疗影像、病例数据，辅助医生更准确诊断疾病，提高早期发现率

  依据患者基因数据、生活习惯等，提供个性化医疗方案推荐与健康管理服务

药物研发加速

- 分析药物分子结构、药理作用数据，加速研发进程，降低研发成本。建立AI训练及考核评价系统，助力医疗人员训练考核，提升专业能力

医疗康养服务

- 提供中医养生知识AI问答，根据患者信息推荐亚健康调理方案。针对特殊人群，如儿童、孕妇，智能回答常见问题，给出特定调理方案

金融行业应用

场景

智能营销助手	智能办公助手	智能客服助手	智能投研辅助	智能风控辅助
营销话术生成	会议/日志	坐席辅助	摘要生成	财务异常分析
营销物料生成	智能助手	聊天机器人	风险传导	舞弊动机识别
实时数字人客服	客服/培训	智能外呼	观点提取	违约风险分析

- 分析大量金融数据、市场动态及用户行为，构建精确风险模型，及时发现潜在风险
- 提升金融机构风险防控能力，避免风险扩大，保障金融稳定

风险评估与预测

- 为投资者提供智能投资建议与资产配置方案，根据市场变化实时调整策略
- 提升投资收益，推动金融行业智能化转型，满足客户多样化需求

投资决策优化

- 在反欺诈方面，利用模型智能分析能力，快速识别异常交易行为，保障金融交易安全
- 提升服务效率，优化客户体验，增强金融机构竞争力

客户体验提升

通信行业应用

智能客服服务

1

- 借助DeepSeek语言理解与生成能力，快速准确回答用户咨询，解决问题，提升服务效率
  提高用户体验，降低客服成本，增强通信企业服务能力

网络优化与运维

2

- 分析海量网络数据，预测网络拥塞点，提前调配资源，保障网络稳定高速运行
  构建智能运维系统，实现网络故障自动预测、定位和修复，提高运维效率

通信行业

业务创新拓展

3

- 助力开发个性化通信套餐推荐、智能通信内容创作等新型业务
  满足用户多样化需求，推动通信行业创新发展，提升企业竞争力

DeepSeek将推动通信行业向智能化、个性化、高效化方向发展，提升服务质量和运营效率

DeepSeek赋能制造行业

预警故障、优化生产、风险评估

行业痛点		
• 设备故障预测不准 • 生产排程优化困难 • 供应链异常响应慢	预测性维护	DeepSeek分析设备传感器数据，提前预警故障，减少停机时间。
	多目标排产优化	通过深度学习算法，DeepSeek优化生产排程，提高库存周转率。
	供应链风险画像	DeepSeek构建供应商风险评估模型，异常响应时间缩短至分钟级。

中国电信 CHINA TELECOM | 5G

感谢各位领导聆听
谢谢！