第 1 部分

源神 DeepSeek

DeepSeek 以开源的大型语言模型迅速崛起，引起全球关注。其最新模型 DeepSeek-R1 在性能上可比肩顶尖 AI 模型，却以极低成本和超短研发周期完成，令业界震惊。DeepSeek 的核心优势在于高性能、开源和低成本，更令人瞩目的是其成本优势，低成本配合高性能，使得以前因成本高昂而无力涉足大模型的中小企业如今也能负担 AI 应用。这一“低成本高性能”的范式突破，被誉为 AI 领域的“斯普特尼克时刻”。

第 1 章

DeepSeek 崛起：重塑全球 AI 版图

DeepSeek-R1 的问世堪称一个里程碑，它标志着开源力量在全球 AI 竞赛格局中的一次重大突破，为中国在全球大模型领域赢得了重要地位。这一成果不仅成功打破了国际科技巨头对 AI 核心技术的长期垄断，还以开放共享的模式，为全球 AI 开发者提供了崭新的技术平台，进一步推动人工智能向普惠化、民主化方向发展，同时加速了全球 AI 创新生态的重塑。

1.1　DeepSeek 系列横空出世

DeepSeek 凭借创新的动态 MoE 架构与强化学习驱动的推理能力，在短时间内迅速获得全球关注，打破了“大模型必须依赖海量数据和算力”的行业认知。随着 DeepSeek 引领的开源模式加速全球开发者生态的构建，中国 AI 企业正从追随者迈向引领者，与美国实验室的封闭模式形成鲜明对比，重塑全球 AI 竞争格局。

1. 主要模型发布时间线

2024 年 11 月 20 日，DeepSeek 在 Twitter 低调发布 DeepSeek–R1 Lite Preview，其推理性能首次引发技术圈关注。尽管在 AIME 数学竞赛中的准确率超越 OpenAI 的 o1–preview，并实时展示了数万字符的思维链过程，但彼时外界更聚焦于 o1 模型，DeepSeek–R1 Lite Preview 的讨论仅限于学术小圈层，甚至被误认为“中国版 GPT–4o 的模仿品”。这种“技术惊艳却无人识”的落差，为后续的爆发埋下伏笔。

转折点出现在 2024 年 12 月 26 日 DeepSeek–V3 的发布。这款采用动态 MoE 架构的模型，通过 37B 参数实现推理速度提升 3 倍，训练成本仅为同类模型的 5% ～ 10%。更关键的是，其技术报告首次公开强化学习驱动的推理能力训练方法，彻底打破“大模型必须依赖海量数据和算力堆砌”的行业认知。安德烈·卡帕斯（Andrej Karpathy）在 Twitter 上盛赞：“这是我所见过的最详细的技术报告，DeepSeek 正在重新定义 AI 效率”，标志着其正式破圈进入学术界视野。

2025 年 1 月 15 日，DeepSeek 应用正式上线，内置 DeepThink 模式。然而，由于当时全球关注特朗普“登基”事件，政治博弈成为焦点，DeepSeek 未能引起广泛注意。直到 2025 年 1 月 20 日，DeepSeek–R1 版本发布，论文与模型权重同步开源。该版本采用纯强化学习训练路径，并利用 2000 块 H800 芯片实现了顶级性能的颠覆性创新，迅速引发全球关注。

根据星摩尔（SimilarWeb）的数据，2025 年 1 月 27 日，DeepSeek 官方网站的日访问量达到 4900 万次，比前一周增长了 614%，甚至一度超过了已推出近两年的谷歌 Gemini 聊天 AI 网站的用户数。在 App 应用方面，DeepSeek 发布一周内累计下载量达 160 万次，覆盖美国、英国、加拿大、新加坡、澳大利亚等主要市场，跃居中美两国 App Store 免费榜首。

在短短 20 天内，DeepSeek 移动端应用的日活跃用户数突破了 2000 万，充分展示了用户对该开源模型的强劲需求。这一现象推动了中芯国际等中概股逆势上涨，

形成了鲜明的“东升西落”格局，彰显了其广泛的市场影响力。DeepSeek 模型发布的主要时间线如图 1–1 所示。

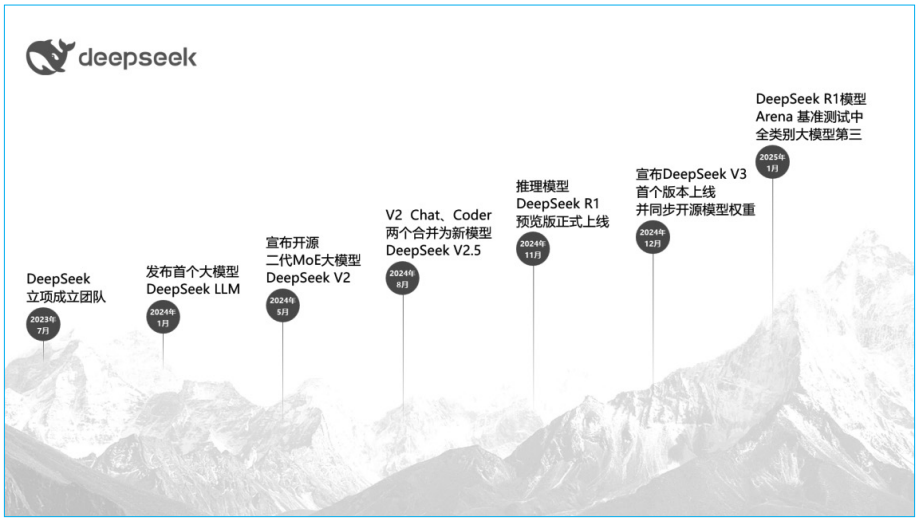

图 1–1　DeepSeek 模型发布主要时间线

2. AI 风暴带来深远影响

这场 AI 风暴的冲击波远超技术层面。华尔街分析师指出，DeepSeek–R1 以 4 元 / 百万 token 的输入成本，直接动摇了英伟达万亿市值的算力霸权。2025 年 1 月 27 日美股暴跌当天，谷歌搜索“DeepSeek”关键词的 70% 查询来自华盛顿特区 IP。Meta 工程师在匿名社区直言：“我们正在疯狂分析代码，试图复现他们的训练方法”，而 OpenAI 前工程师则在播客中感叹：“这可能是中国 AI 的‘ChatGPT 时刻’”。

更深远的影响体现在产业生态上。DeepSeek–R1 通过蒸馏技术将模型参数规模压缩至 7B~8B 级别，开发者仅需单张消费级显卡即可完成本地部署。这种“技术民主化”趋势，为中小企业和开发者提供了“万元级”AI 普惠化解决方案。当前，DeepSeek 日活跃用户突破 3000 万，7 天获客 1 亿，增速超越 ChatGPT，创下了新的纪录。其 API 服务更推出“错峰定价”策略：夜间时段调用成本直降 75%，DeepSeek–V3 与 DeepSeek–R1 输入同价至 1 元 / 百万 token。这种“技术普惠 + 商业创新”的组合拳，正在书写 AI 发展史的新篇章。

这场技术民主化革命，使得 DeepSeek–V3 训练成本降至 558 万美元，仅为 GPT–4o 的 1/18。当其开源代码库在 GitHub 上星标破万，亚马逊、微软等巨头排队接入时，这场由东方公司发起的革命，正在改写全球 AI 权力格局。正如图灵奖

得主杨立昆（Yann LeCun）所言：“DeepSeek 的成功证明了开源模型正在超越专有模型”。

3. 从 AI 追随者到引领者

中国 AI 企业的崛起正重新定义全球竞争格局。以 DeepSeek 为代表的中国 AI 实验室通过持续突破，完成了从技术追随者到创新引领者的跨越。2025 年年初发布的 DeepSeek–R1 推理模型，其基准测试表现已逼近美国 OpenAI 的 o1 性能水平，标志着中美模型智能差距从代际级缩短至毫厘之间，如图 1–2 所示。这场“伟大的追赶”背后，是中国特有的发展路径。

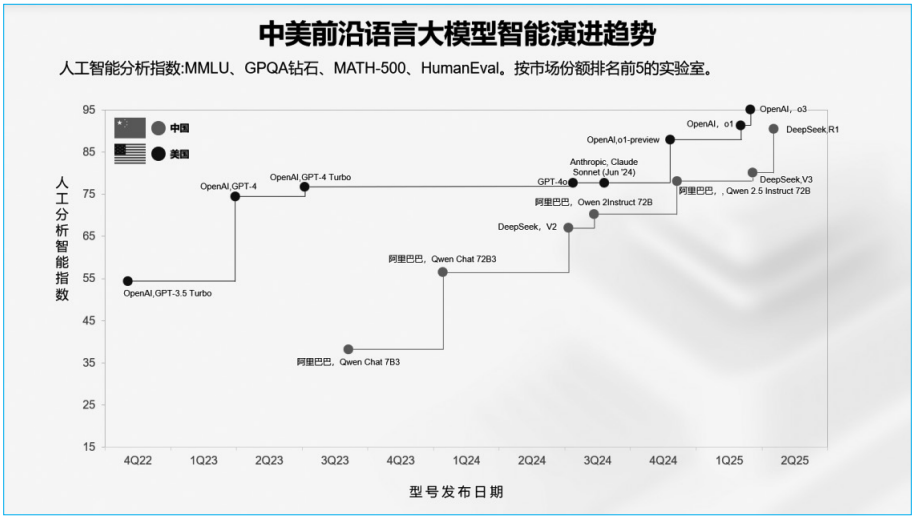

图 1–2　中美前沿模型竞争

中国的人工智能领域正经历一场前所未有的变革，形成了“1 + 6 + X”的多元化创新生态。这一生态涵盖了以深度求索（DeepSeek）为代表的先锋企业，以及智谱 AI、Minimax、百川智能、零一万物、阶跃星辰、月之暗面等“AI 六小龙”，还有阿里、腾讯、百度、字节、华为等传统科技巨头。

中国企业通过开放模型权重的策略，在全球竞争中脱颖而出。这种开放模式吸引了全球开发者参与生态共建，降低了技术门槛，加速了创新迭代。与此形成鲜明对比的是，美国实验室依赖封闭生态（如 OpenAI 的 o3 系列闭源模型），同时继续借助英伟达 GPU 等硬件优势维持技术壁垒。这预示着未来竞争将呈现“软件开源突围”与“硬件封锁博弈”并存的复杂态势。

1.2 DeepSeek-R1 的重要意义

DeepSeek-R1 作为纯国产 AI 大模型，通过自主研发的“专家协同架构”与强化学习创新技术，实现了参数效率与推理能力的突破。其深度软硬件协同优化使训练成本仅为国际同级别模型的几十分之一，并以“高性能、低成本”普惠化，开创了国产 AI 技术自主创新与商业效率双突破的新范式。

1. 纯国产与技术自主创新

DeepSeek-R1 取得巨大成功的关键在于坚持纯国产路线，完全由中国团队自主开发，没有依赖西方开源模型改造。其前身 DeepSeek-V3 发布时性能已与国际一流的闭源模型不相上下，而 DeepSeek-R1 进一步强化了推理能力，实现了真正意义上的“中国制造”。

团队创新性地设计了一种特殊架构（Mixture-of-Experts, MoE 架构），通过整合 256 个领域专家与 1 个共享专家，配合智能动态路由机制，实现仅激活 8 个最优专家的高效计算模式，大幅减少了计算量，达到参数规模与计算效率的完美平衡。团队还自主研发了一种无辅助损失的负载均衡策略算法（Auxiliary-Loss-Free Load Balancing，负载均衡策略），确保每个专家网络均衡地分担工作，使整体计算效率显著提高。此外，DeepSeek-R1 还加入了强化学习的创新方法（GRPO，群体相对策略优化），让模型能自我评估、多方案竞争选择最佳方案，进一步提升了复杂问题的解决能力。

在硬件方面，团队通过直接使用 NVIDIA 的底层指令集（Parallel Thread Execution，PTX）对 GPU 通信和计算进行了微优化，这一举措被媒体戏称为“绕过 CUDA 垄断”，最大程度地榨取了硬件性能。通过深度优化软件和硬件协作，性能接近国际顶级水平。这些努力不仅证明了中国 AI 技术的实力，也探索出一条在有限资源条件下，通过算法创新突破瓶颈的有效路径。

2. 高效率地训练与部署

传统的大模型训练成本高昂，DeepSeek-R1 则通过一系列创新措施显著降低了成本与资源消耗。尽管模型规模极大，但每次推理仅用到很少一部分参数，大大提高了计算效率，使在有限硬件资源条件下也能完成训练。DeepSeek 官方声称 DeepSeek-R1 的总训练开销在 558 万美元左右，与 GPT-4 等模型相比便宜了几十倍。DeepSeek-R1 充分利用了混合精度计算的红利。它是业界最早大规模采用英伟达低

精度训练（8-bit Floating Point，FP8）的大模型之一。

DeepSeek-R1 通过 MoE 架构 + 并行算法 + 低精度训练的组合拳，成功突破了大模型训练和部署的效率瓶颈。在算力受限的条件下做出了接近“地表最强”性能的模型，实现了高性能和高成本效益的兼得。这一成就向业界证明：通过技术创新，我们可以走出一条“高效大模型”的新路，为 AI 大模型的可持续发展提供了宝贵经验。

3. 性能比肩国际顶级水平

DeepSeek-R1 在实际测试中，综合表现已经达到国际顶级闭源模型的水平，甚至在某些方面表现更为突出。在数学推理基准 MATH-500、AIME 竞赛题等高难度任务上，DeepSeek-R1 的得分超过了 OpenAI 的 ChatGPT-4 和 Anthropic 最新的 Claude 3.5（Sonnet 版）。有业内人士直言：“DeepSeek-R1 在聪明程度上明显强于 Claude 3.5、OpenAI o1-Pro，甚至优于谷歌的 Gemini”。

更难能可贵的是，DeepSeek-R1 在实现高性能的同时保持了稳定输出和可靠结构化响应。这使其不仅擅长学术测试，在实际应用中也足以媲美 GPT-4、Claude 等的表现。综合来看，DeepSeek-R1 已经站上了当前 AI 大模型性能的第一梯队，其在推理深度和问题解决方面的领先性为行业树立了新标杆。随着 DeepSeek-R1 不断优化迭代，有理由相信它在更多基准上全面超越 GPT-4、Claude 3、Gemini 等主流模型只是时间问题。

4. 大幅降低的使用成本

DeepSeek-R1 不仅技术强大，在商业层面也祭出了“价格屠夫”的策略。与 OpenAI、Anthropic 等顶尖 AI 服务高昂的 API 费用相比，DeepSeek-R1 的使用成本可谓低到难以置信。根据官方公布的数据，DeepSeek-R1 输出每 1M 字符的费用仅相当于 o1 的 1/30。这种价格优势让高端 AI 技术首次实现了普惠化，即使中小企业也能轻松负担。

更重要的是，这种低价策略并非以牺牲利润为代价，而是通过极高的技术效率实现了成本与利润的双赢。据官方测算，即使以如此低的价格提供服务，DeepSeek 团队依然能够维持高达 545% 的利润率。

这种模式不仅颠覆了过去 AI 服务高价的传统，也迫使国际巨头重新审视自己的定价策略。DeepSeek-R1 的出现标志着 AI 服务开始进入全民可负担的新阶段，使“高性能、低成本”的 AI 成为现实。

5. 推动生态多元化发展

DeepSeek-R1 作为开源大模型，以接近顶尖封闭模型的性能提升了行业标杆，推动大模型服务（Metal as a Service，MaaS）平台生态进入新阶段。它证明了算法优化（如纯强化学习后训练）可以带来突破，促使 OpenAI、Anthropic 等巨头调整策略，并加速国内初创企业的优胜劣汰。

DeepSeek-R1 已成为 2025 年以来 MaaS 平台的“共同语言”，云厂商纷纷集成其模型，在模型供应同质化的背景下，竞争焦点正逐步转向并发能力、响应延迟等服务质量的提升。同时，它对初创企业形成竞争压力的同时，也带来合作机遇，资本方加速开源派与自研派的融合，推动智谱等公司发布新一代开源模型。

作为开源生态的中流砥柱，DeepSeek-R1 采用 MIT 许可，激活了社区创新，催生出多个蒸馏版本和融合模型，增强了“开源可行”的信心，促进生态去中心化。官方团队在 GitHub、知乎等平台分享优化经验，提升行业技术水平。

DeepSeek-R1 还重塑了 MaaS 价值链：模型商品化趋势增强，云厂商更重视附加服务，初创企业需探索模型之外的价值，算力和算法的重要性平衡也发生变化。用户得以在不同平台间灵活选择最优方案，降低迁移成本。DeepSeek-R1 承上启下，推动 MaaS 生态走向更加开放、多元、以技术创新和性价比驱动的新阶段，并将在未来迭代中持续发挥关键作用。

1.3 DeepSeek-R1 定义推理新纪元

推理型大模型的出现标志着 AI 发展的重要变革：从单纯追求模型规模到更强调模型的“聪明程度”。DeepSeek-R1 的成功展示出，通过强化学习、自我反思和知识蒸馏，AI 模型能够持续自我进化并表现出更高的推理和决策水平。未来 AI 领域的竞争核心将不再是规模大小，而是推理能力和实际应用的有效结合。

1.3.1 推理大模型的由来

2024 年 9 月 12 日，OpenAI 官方正式发布 o1-preview 推理大模型，标志着人工智能认知能力的重大突破。该模型引入了“推理时计算（test-time compute）”的全新理念，即在推理阶段（模型生成最终结果的过程中）动态投入更多计算资源，而不仅仅依赖于预训练或后训练阶段的优化。

这一创新使模型具备更强的内部思考能力，能够评估多个潜在答案、进行深度规划，并在输出最终结果前进行自我反思。典型应用包括“思维链推理（Chain-of-

Thought Reasoning）”和“Wait 注入（Wait–and–Inject）”技术，它们拓展了模型的推理空间，大幅提升逻辑推理和问题求解能力。

推理型大模型被誉为天生的战略家，因其基于数据分析和逻辑推理，能够自主学习、推理和决策。这类模型擅长从已知信息中挖掘潜在规律，并广泛应用于数学、代码、逻辑推理等高认知任务领域。图 1–3 直观地展示了该模型在这些领域的卓越表现。

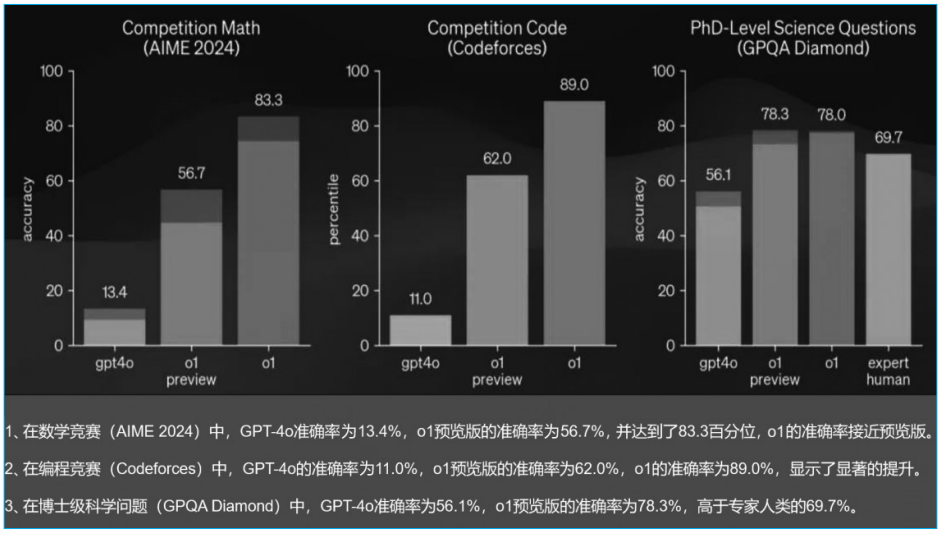

图 1–3　两类模型能力对比

这一转变带来的影响是深远的。在过去，如果模型无法在 200ms 内生成响应，几乎等同于失败。然而，到了 2025 年，经验丰富的搜索开发者和检索增强生成（Retrieval–Augmented Generation，RAG）工程师们已逐步将精确率和召回率视为首要优化目标，而非单纯追求响应速度。如今，大多数用户已习惯看到系统“努力思考”（<thinking>），只要最终结果足够优质，等待时间便不再是主要问题。

这一趋势与过去对生成式模型的需求形成了鲜明对比，也在深刻改变着用户体验。用户正逐步接受“延迟满足”的交互模式，即通过更长的等待时间换取更高质量、更具实用性的回答。DeepSeek–R1 进一步强化了这一体验，使用户，即便不自觉地，也逐渐适应了更长的响应时间。

OpenAI 宣称，o1 系列大模型在推理（Reasoning）能力方面相较于当前主流模型（如 GPT–4o）有了显著提升。这一进步得益于一种新的 AI 训练方法，强调“思维链推理”过程以及强化学习（Reinforcement Learning，RL），最终使得 o1 系列在

数学、逻辑与推理方面取得了突破性进展。因此，人们开始将具备强大推理能力的 AI 统称为推理大模型（Reasoning AI）。除 DeepSeek–R1 外，市场上的主流推理大模型还包括 ChatGPT（o1/o3）、Gemini 2.0、Grok 3、Kimi 1.5，如表 1–1 所示。

表 1–1　主流推理大模型

模型名称	开发公司	特点
OpenAI o1 2024 年 12 月	OpenAI	OpenAI 推出的首个具备推理能力的模型，能够在回答前进行深入思考，提升在科学、编程等领域的表现
OpenAI o3–mini 2025 年 1 月	OpenAI	o3 的轻量版，提供高效且经济的推理能力，特别在科学、数学和编程等领域表现出色
Gemini 2.0 2024 年 12 月	谷歌	谷歌的高级 AI 模型，具备多模态输入 / 输出能力，性能是前代 1.5 Pro 的两倍，支持复杂任务的处理
Grok 3 2025 年 2 月	xAI	追求超越传统 AI 的推理能力，强调逻辑推理、深度理解和动态决策，不仅仅是数据驱动的统计学习
Kimi 1.5 2024 年 12 月	月之暗面	原生支持端到端图像理解和思维链技术，能力扩展到数学之外的更多领域

1.3.2　思维链展示过程

DeepSeek–R1 的突破在于其成为首个能够展示自身思维链的 AI 推理模型。传统大型语言模型往往仅给出最终答案，内部推理过程对用户而言是“黑箱”。DeepSeek–R1 通过引入显式的思维链（Chain–of–Thought，CoT）生成机制，使模型在回答问题时先产出逐步推理步骤，再得出最终结论。在训练过程中，DeepSeek–R1 被要求以特殊标记输出思考过程，从而学会自主生成多步推理链。在实际应用中，用户可要求 DeepSeek–R1 展示其思路，清晰地看到逻辑推演过程及最终答案，如图 1–4 所示。

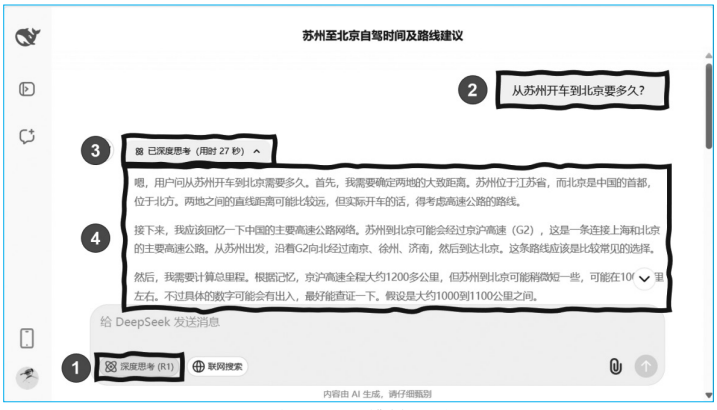

图 1–4　思维链展示

用户单击“深度思考（R1）”按钮❶后，输入提示词“从苏州开车到北京要多久？”❷，此时系统将调用 DeepSeek-R1 进行推理。首先，界面会显示“思考中”❸，在经历响应延迟阶段后，模型开始输出思维链。当思维链输出完成后❹，界面显示“已深度思考（用时 27 秒）”，请注意，此处的 27 秒并不包括最初的响应延迟。当思考完成后，模型输出最终响应。值得注意的是，每次输出的思维链内容都会有所不同。

以上示例，也说明了思维链展示的价值在于以下几个方面。

1. 透明性与可解释性的提升

这一能力首先增强了 AI 的透明度和可解释性。由于思维链清晰可见，用户能够理解模型如何得出某个结论，从而更容易信任其决策。尤其在金融决策、医疗诊断等高风险场景，这种可解释性至关重要。

2. 推理深度与自我反思机制

思维链有助于提升模型的推理深度。DeepSeek-R1 在推理过程中能够自我检查、反思，并“看到”前序步骤中的错误，从而尝试修正。实验证明，随着训练的深入，DeepSeek-R1 逐渐涌现出长链条推理能力，甚至展现出类似人类“豁然开朗”的“aha 时刻”，能够自主发现并纠正推理中的问题。这种自我反思机制使解题过程更加严谨，避免了“一步到位”式回答中常见的荒谬错误。

3. AI 推理范式的革新

DeepSeek-R1 开创性地将 AI 的思维过程从幕后呈现至前台，不仅让 AI 具备“会思考”的能力，更让用户能够理解 AI 在思考什么。这一能力大幅提升了 AI 系统的可信度和推理能力，使其在复杂任务上展现出优于以往模型的表现。DeepSeek-R1 的成功验证了“让 AI 自主展示思路”这一范式的可行性，为行业树立了新的标杆。

1.3.3　自我进化的秘密

你是否有过这样的体验？当老师问“为什么 1 + 1 = 2？”，我们不仅会直接回答“因为数学规则”，还会补充“因为一个苹果加一个苹果是两个苹果”。这种“解题过程”的思维方式，被称为推理能力（Reasoning），也是科学家一直想让机器学会的技能。

2025 年，DeepSeek-R1 的问世，终结了语言模型只能“猜答案”的时代——它通过一种神奇的方法，让模型自己从训练数据中“悟出”推理过程，就像人类一样边想边改进。通过以下三个步骤，打造“自我进化”的推理大脑。

1. 第一步：用“规则 + 强化学习”激发自我修正能力

DeepSeek–R1 的关键突破在于，它完全依靠强化学习（RL），让模型自己从错误中学习。具体来说，训练过程分为以下三步。

（1）定义评分规则：模型输出的内容必须满足两个条件——答案正确、推理过程符合 <think>…</think><answer>…</answer> 的格式。

（2）自我对弈与反馈：模型在生成答案时，会自动尝试多种可能性（类似决策树的“分岔路径”），并根据结果自动调整策略。例如，在解题时，若某个步骤导致错误，模型会自动“回溯”并修正思维过程。

（3）优化与迭代：通过数十亿次“自我对弈”，模型最终学会“如何高效且正确地思考”，甚至能自发出类似人类的“反思”行为（如发现错误后重新检查）。

2. 第二步：从“天马行空”到“清晰可读”，冷启动与格式优化

尽管 R1–Zero（纯强化学习训练版本）展现了惊人的推理能力，但其生成的内容存在中英文混杂、格式混乱等问题，难以在实际应用中推广。为此，DeepSeek–R1 引入了冷启动（Cold Start）技术，通过以下措施加以改进。

- **优秀范例示范：**从 DeepSeek–R1–Zero 的输出中筛选出上千条高质量的推理过程，作为模型学习的范本，这些范例经过人工后处理，确保格式规范、推理清晰，帮助模型在初期阶段建立良好的输出习惯。
- **格式强制约束：**在强化学习训练中加入“语言一致性”奖励，鼓励模型使用统一的语言和格式，例如，计算推理过程中目标语言的比例，避免出现中英文混杂的情况，确保逻辑步骤清晰明了。

通过这些改进，模型生成的推理过程从“天马行空”转变为“结构清晰”，甚至能够像数学老师一样逐步推导，提升了实际应用的可行性和用户体验。

3. 第三步：小模型通过“模仿学习”实现逆袭

然而，RL 训练需要庞大的计算资源，直接用于小模型（如 320 亿参数的 Qwen–32B）时，效果往往大打折扣。为了解决这一问题，DeepSeek 采用了知识蒸馏（Knowledge Distillation）方案：如同师父带徒弟一般，利用 DeepSeek–R1 生成的高质量推理数据训练小模型，使其模仿“大模型师父”的思维模式。这一方法成效显著，经过知识蒸馏训练后，Qwen–32B 在数学题上的表现从纯 RL 训练的 58 分跃升至 74 分，甚至超过了部分更大规模的模型。

目前，DeepSeek 已经推出多个蒸馏版本，包括基于 Qwen 和 LLaMA 的 1.5B 到

70B 模型，大幅降低了推理能力的应用门槛。未来，大模型的竞争或许不再只是“谁更大”，而是“谁更聪明”——但真正的决胜关键，仍在于如何将这些技术与实际场景结合，并通过用户数据持续优化。

正如 DeepSeek 所展示的那样，AI 的进化永远充满惊喜，而我们都是这场革命的见证者。

1.4　DeepSeek 生态：国内外全景

DeepSeek 的出现重塑了全球 AI 竞赛格局，推动技术从“参数竞赛”转向“成本控制与场景适配”。DeepSeek 在激发良性技术竞赛的同时，以开源合作的方式为国内外 AI 生态注入活力——顶尖开发者们开始仔细分析和借鉴 DeepSeek 高效训练、低成本运行的技巧。

1.4.1　国际大厂的接入情况

因为成本大幅降低，算力门槛下降，各行各业竞相尝试将 AI 大模型融入自身业务。过去被高昂算力和数据要求阻碍的应用，如今因 DeepSeek 的“算力平权”而得到落地良机。DeepSeek 证明了有限算力也能训练出强大模型，这增强了企业对自主可控 AI 技术的信心，并可能加速 AI 领域“东升西降”的竞争态势。

DeepSeek 以技术和模式创新在全球 AI 版图中占据了举足轻重的位置：一方面引领了 AI 技术开放普惠的新潮流，另一方面倒逼国际巨头加速创新，全面推动了产业链上下游的革命性变革。以下是国际主要云厂商接入 DeepSeek 的时间线及特点。

1. 微软（2025 年 1 月）

作为首批响应的国际巨头之一，微软 CEO 萨提亚·纳德拉（Satya Nadella）在 2025 年 1 月 29 日的财报电话会上宣布，已将 DeepSeek-R1 模型接入微软 Azure AI 平台的企业级服务（Azure AI Foundry），并通过 GitHub 向开发者提供访问。

微软选择接入 DeepSeek 的出发点在于丰富其 Azure 云 AI 生态，确保客户能够获得最新最强的 AI 能力，同时分散对单一供应商（OpenAI）的依赖。在 OpenAI 的 GPT 系列之外增加 DeepSeek，可以满足更多元的核心需求：一方面，Azure 企业客户希望使用高质量对话与推理能力模型改进业务应用；另一方面，微软也希望借 DeepSeek 超低的使用成本，提供更具价格竞争力的 AI 服务。

DeepSeek-R1 开源且性能领先，这些技术特点契合了微软的需求：DeepSeek-R1

拥有超长上下文和强推理能力，可用于企业文档分析、代码助手等复杂场景；同时 DeepSeek 模型的高效架构使其在 Azure 上部署算力要求相对较低、调用成本低廉，有助于微软以更低的价格为客户提供大模型服务。微软 Azure 接入 DeepSeek 后，开发者能够零门槛调用这一先进模型，提高了 Azure 平台的吸引力和竞争力。

2. 谷歌（2025 年 2 月）

与其说谷歌“接入”DeepSeek，不如说是被 DeepSeek 逼出了杀手锏。DeepSeek–R1 的横空出世在 2025 年年初抢尽风头：它与谷歌自研的推理模型几乎同时在 2024 年 12 月发布，但显然后者的关注度被 DeepSeek–R1 盖过。尤其是在春节前夕，DeepSeek 一度登顶全球应用商店下载榜，令谷歌深感压力。谷歌的应对措施是在 2025 年 2 月 5 日紧急发布新一代 Gemini 2.0 系列模型，包括 Flash、Flash–Lite 和 Pro 三个版本。其出发点在于通过自研更强模型来重夺业界领导地位，避免用户和开发者被 DeepSeek 的风潮吸引走。

Gemini 2.0 系列主打更高性能和更优成本。据谷歌介绍，Flash 模型提供了更高的速率和性价比，Pro 模型拥有高达 200 万 token 的上下文窗口并可调用谷歌搜索和代码执行工具，专攻复杂推理。可以看出，谷歌在核心需求上与 DeepSeek 展开正面竞争——提供更大的上下文、更强的多模态推理以及更低的使用成本，这些都是 DeepSeek 引领的方向。

DeepSeek 模型的成功迫使谷歌加快了 Gemini 的升级发布节奏，甚至谷歌 DeepMind CEO 德米斯·哈萨比斯（Demis Hassabis）也公开评价称 DeepSeek 是“中国最好的作品”，在短时间低成本训练方面表现惊人。不过谷歌同时淡化了 DeepSeek 的技术创新程度，认为其“使用的都是已知技术，很多炒作有些夸大”。谷歌强调自家最新版 Gemini 在效率上更胜一筹。总的来看，谷歌并未直接整合 DeepSeek 到自家产品，而是选择了技术跟进和迭代超越的策略：以更强大的自研模型来回应挑战。

这场你追我赶也凸显了 DeepSeek 对谷歌的冲击之大——逼得这位搜索巨头“卷”了起来，加速开放其最强 AI 模型在云服务（Vertex AI）和消费应用中的使用。谷歌的快速反击证明了 DeepSeek 在国际 AI 竞赛中已经成为不可忽视的力量，直接推动了巨头产品路线的调整。

3. Meta（2025 年 1 月）

DeepSeek 的出现同样令 Meta 公司如坐针毡。2025 年 1 月 27 日前后，有媒体披露 Meta 的生成式 AI 团队因 DeepSeek 的震撼表现而陷入恐慌，团队主管公开表达了

对自家模型可能落后的担忧。据信息时报（The Information）报道，Meta 的 AI 研究人员甚至火速成立了 4 个“战情室”来研究 DeepSeek。这一举措大约发生在 2025 年 1 月下旬。可见 DeepSeek-R1 发布仅一周内就引发了 Meta 的高度紧急响应。Meta 此举的出发点在于防止在新一轮 AI 竞赛中掉队。作为开源大模型 LLaMA 系列的推出者，Meta 原本在开源社区具有优势地位，但 DeepSeek 的横空出世以更强性能抢走了风头，也让 Meta 意识到自身在训练成本和效率方面的不足。

因此，战情室的工作聚焦在 Meta 的核心需求上：①寻找降低大模型训练和推理成本的方法，以期追赶 DeepSeek 实现的效率奇迹；②分析 DeepSeek 究竟用了哪些数据和技巧训练模型，以借鉴其高效训练范式；③考虑对自家模型架构进行调整，如是否采用多模型专精的路线。报道指出，Meta 高层正考虑推出类似 DeepSeek 的 LLaMA 版本，即由多个专门擅长不同任务的模型组合而成。

DeepSeek 的做法启发了 Meta 思考“一个模型未必包打天下，不妨训练多个各有所长的模型”。在技术特点上，DeepSeek 对 Meta 最有借鉴意义的是其低成本高效训练（仅用 OpenAI 十分之一的开销）和多专家分工的模型策略，这恰是 Meta 提升 LLaMA 系列所需的方向。

可以说，Meta 没有直接将 DeepSeek 嵌入自家产品，而是选择研究学习和策略调整：通过内部消化 DeepSeek 的成功经验，来升级自身开源模型，以期在即将到来的竞赛中重新夺回优势。这种反应也印证了 DeepSeek 在国际 AI 舞台上的影响力——连拥有最顶尖 AI 研究团队之一的 Meta 都不得不“闭门苦练内功”，以应对这匹黑马带来的挑战。

4. 亚马逊（2025 年 1 月）

2025 年 1 月 31 日，亚马逊宣布其云服务 AWS 已上线 DeepSeek-R1 模型，供用户直接调用。AWS 接入的动机在于巩固其云 AI 领先地位，不让微软独享这一热门模型。

面对企业用户对生成式 AI 日益增长的需求，AWS 需要满足核心需求：为客户提供最前沿的 AI 能力并简化部署。DeepSeek 的特点正契合 AWS 所求——作为开源模型，客户可在 AWS 上自由调整和私有部署，符合企业对数据隐私和定制化的要求；同时，DeepSeek-R1 被公认为当时最先进的大模型之一，语言处理效果卓越。

AWS 将其纳入模型库后，开发者通过 AWS 即可低门槛获取高性能的 DeepSeek 服务，从而提升了云平台对开发者和企业的吸引力。亚马逊还看重 DeepSeek 低算力占用的优势，这有助于 AWS 优化其算力资源利用率，并通过更低的调用成本吸引大

批客户迁移到自己的云上。

5. 英伟达（2025 年 1 月）

2025 年 1 月 31 日，英伟达在官网宣布，其 NVIDIA NIM 平台现已支持调用 DeepSeek–R1 模型。作为 AI 计算领域的领导者，英伟达的出发点是展示自家硬件对新兴顶级模型的良好支持，并推动 GPU 算力需求进一步增长。

其核心需求在于确保 DeepSeek 这样的现象级模型能在英伟达的 GPU 和软件栈上高效运行，从而巩固英伟达在 AI 基础设施上的统治力。DeepSeek–R1 的技术特点对于英伟达而言具有双重意义：一方面，DeepSeek 团队以远低于以往顶级模型的算力完成训练，这对英伟达既是冲击也是机遇——它促使英伟达优化软硬件以支持这种高效模型，吸引开发者继续选择 GPU 作为主要运行环境；另一方面，DeepSeek 高达 6710 亿参数且含多模态模型的规模，依然需要强大 GPU 集群才能充分发挥，其开源发布将刺激众多企业部署实验，这无形中扩大了英伟达 GPU 的市场需求。

通过 NIM 平台集成 DeepSeek，英伟达展示了自家 GPU 对该模型的兼容与性能优势，为客户提供开箱即用的 DeepSeek 优化推理服务，进一步稳固了“用最强 GPU 跑最强模型”的市场形象。

6. 英特尔（2025 年 2 月）

2025 年 1 月 31 日，英特尔迅速完成了对 DeepSeek 模型的适配，展示了其 AI 硬件生态对最新大模型的支持，避免在这波 AI 热潮中被边缘化。英特尔的核心需求在于优化自家 GPU、CPU 及 AI 加速器，以高效运行 DeepSeek，提高软硬件结合的 AI 算力性价比，向市场证明除了英伟达 GPU 外，还有其他可靠选择。

DeepSeek 开源提供模型权重，这一技术特点方便英特尔工程团队快速移植和调优。例如，英特尔将 DeepSeek 部署在其至强（Xeon）CPU 和 Gaudi AI 加速器平台上进行测试。DeepSeek 模型本身对硬件依赖度低、可伸缩性好，证明了即使不完全依赖最尖端 GPU 也能取得顶级 AI 表现。这为英特尔等厂商提供了契机——通过支持 DeepSeek，它们可以宣传“在我们的硬件上，同样能以更低成本跑出一流 AI 效果”，以此吸引云服务商和企业采用自家芯片方案。

1.4.2 国内大厂的接入情况

海外巨头牵手 DeepSeek 的消息传来之际，中国本土的科技企业也迅速行动起来。春节假期期间（2025 年 1 月底至 2 月初），国内各大云计算厂商加班加点完成了对 DeepSeek 模型的适配上架。包括华为云、腾讯云、阿里云、百度智能云、字

节跳动火山引擎、京东云在内的主流云服务平台都在 2025 年 2 月上旬相继宣布支持 DeepSeek 系列模型。

这些云厂商的出发点非常一致：借助 DeepSeek 的热度和技术优势，完善自身 AI 算力生态，吸引开发者和企业客户留在或迁移到自己的云上。它们的核心需求在于快速提供开箱即用的大模型服务，降低用户使用门槛，从而扩大市场份额。

在这方面，DeepSeek 模型开源且可私有部署的技术特点给予了云厂商极大的便利——相比只能通过 API 云调的封闭模型，DeepSeek 允许云服务商将模型部署在本地算力池并深度优化。由此，各家云厂商纷纷基于各自主推的平台优势对 DeepSeek 进行整合，推出特色服务和优惠策略：有的提供限时免费调用、有的赠送大额 token 额度，以期利用全民关注的“DeepSeek 时刻”笼络用户。以下是国内主要云厂商接入 DeepSeek 的时间线及特点。

1. 阿里巴巴（2025 年 2 月）

阿里云官方于 2025 年 2 月 3 日宣布，其 PAI Model Gallery 模型库已支持一键部署 DeepSeek-R1/V3 模型。阿里云接入的动机在于丰富云上 AI 模型选择，强化其“模型即服务”能力。其核心需求是让企业和开发者零代码即可使用和训练大模型，以提高开发效率。DeepSeek 开源提供完整权重，正好满足这一需求——阿里云通过整合 DeepSeek 模型，用户在其平台上从训练到推理全流程都可一站式完成。

技术上，阿里云针对 DeepSeek 进行了优化适配，使其能够在飞天云基础架构上高效运行，并推出了超低价格方案甚至免费体验，以 DeepSeek 显著的成本优势来吸引客户。阿里云还强调 DeepSeek 模型在多行业的泛用性，这契合阿里云服务众多行业客户的场景需求。

接入 DeepSeek 后阿里云迅速打造出了从模型训练、部署到应用的便捷通道，为开发者和企业提供了更快、更高效的 AI 开发体验。通过拥抱 DeepSeek 的开源生态，阿里云展现了开放兼容的姿态，不仅巩固了国内云市场地位，也为其全球云服务增加了亮点。

2. 百度（2025 年 2 月）

2025 年 2 月 3 日，百度智能云宣布旗下“千帆”平台正式上架 DeepSeek-R1 和 DeepSeek-V3 模型。百度选择接入 DeepSeek 的出发点在于顺应开源大模型浪潮，弥补自身模型在特定方面的不足，提供更加多样化的 AI 服务组合。

百度智能云的核心需求，一是利用 DeepSeek 高推理能力满足客户在复杂问答、

数据分析等场景的需求，二是通过差异化定价抢占市场。DeepSeek 模型的技术特点为百度实现这两点提供了抓手：据发布，百度智能云针对 DeepSeek 推出了业界罕见的低价计费方案，并在一定时期内提供免费调用服务。这背后正是利用 DeepSeek 超低算力成本的优势，让利给用户，从而吸引对价格敏感的中小客户。

与此同时，DeepSeek 模型在自然语言理解和多轮对话上的强大性能使其成为百度自身文心大模型的有力补充。特别值得一提的是，百度还将 DeepSeek 纳入其文心一言生态：文心一言在此期间宣布对全社会免费开放，并计划开源，其搜索引擎、信息流等业务也开始灰度接入 DeepSeek 模型的能力。这说明百度并非仅将 DeepSeek 视为第三方模型，而是积极考虑将其融入自身产品体系，实现优势互补。例如，百度搜索和百家号内容可能通过 DeepSeek 获得更强的生成和分析功能，从而提升用户体验。

3. 腾讯（2025 年 2 月）

腾讯系在这波浪潮中也动作为先。腾讯云在 2025 年 2 月 2 日即实现其云平台全面适配支持 DeepSeek 系列模型，开发者可在腾讯云上直接调用使用。腾讯云接入的出发点在于完善其云 AI 生态和工具链，特别是结合腾讯自身业务（如社交、内容、安全）为客户提供一体化的 AI 解决方案。

其核心需求包括：为企业客户提供高质量对话模型用于客服与运营，为游戏和内容行业客户提供内容生成和审校能力，以及为安全风控场景提供智能分析支持等。DeepSeek–R1/V3 模型优秀的语言理解、逻辑推理性能正好满足这些需求，腾讯云通过优化部署，可以让 DeepSeek 在自家基础设施上实现开箱即用，并借助腾讯丰富的行业知识进一步调优模型效果。

腾讯凭借庞大的 C 端应用生态，将 DeepSeek 接入微信进行灰度测试，引入“AI 搜索”功能，提供“快速问答”和“深度思考”两种模式。DeepSeek–R1 能智能检索公众号、视频号及全网信息，为微信 13 亿月活跃用户带来 AI 搜索能力，助力其打造社交 + 信息 + 服务的一站式智能搜索，覆盖超 10 亿用户。

腾讯的核心目标是强化微信的超级应用地位，提升用户黏性，并依托 DeepSeek–R1 的中文理解与联网检索能力满足即时信息获取需求。此外，DeepSeek 的超长思维链和可私有部署特性，使其在推理问答和数据隐私方面契合微信的技术要求。

通过云端部署 + 终端集成，腾讯实现了从 B 端云服务到 C 端国民应用对 DeepSeek 的全面拥抱。这表明，国内大厂并非只是被动跟随，而是善于将 DeepSeek 的能力融入自身庞大的产品生态中，创造新的价值。

4. 字节跳动（2025 年 2 月）

作为国内内容和互联网领域的巨头，字节跳动也通过其云服务平台火山引擎加入了 DeepSeek 生态 。火山引擎在 2025 年春节前后快速适配了 DeepSeek 模型，向外界提供云上调用。字节跳动的出发点在于借助 DeepSeek 完善其“云 + AI”战略布局，并反哺自身内容业务。

其核心需求一是为广大中小互联网企业提供内容生成、审核类 AI 能力（这是字节跳动擅长的领域），二是将 DeepSeek 的强大通用智能引入自家产品（如今日头条、抖音）的创新。DeepSeek 的技术特点非常契合字节跳动的产品基因：作为开源模型，它允许字节跳动对模型进行本地部署和深度定制，这意味着字节跳动可以针对海量中文内容和推荐场景微调 DeepSeek，训练出更懂国内用户偏好的版本；DeepSeek 在内容生成和对话上性能卓越，能够满足字节跳动系产品在智能创作、智能客服等方面的需求。

DeepSeek 低推理成本的优势也降低了字节跳动在旗下应用中大规模部署 AI 功能的压力（例如，在抖音内给创作者提供 AI 脚本建议，就需要低成本的模型支持）。通过火山引擎开放 DeepSeek 服务，字节跳动向外输送了自身的 AI 能力，并借此吸引更多开发者使用其云服务，从而打造内容生态之外新的增长点。

可以预见，字节跳动未来或将 DeepSeek 能力融入抖音国际版 TikTok 等产品，为全球用户提供更智能的内容创作和交互体验。在这背后，DeepSeek 作为“开放的通用 AI 引擎”正好符合字节跳动崇尚的敏捷试错和规模扩张思路——开源模型让它们可以快速集成、持续优化，以最快速度将 AI 新功能推向亿级用户市场。

5. 华为（2025 年 2 月）

作为国内既有云服务又有芯片硬件布局的科技巨头，华为对 DeepSeek 的态度尤为积极。2025 年 2 月 2 日，华为云联合硅基流动，正式上线基于昇腾云服务的 DeepSeek R1/V3 模型，标志着华为云成为国内首家实现 DeepSeek 全栈国产化部署的运营商级云平台。

华为接入 DeepSeek 的出发点有两方面：其一，发挥自身软硬件协同优势，用国产算力承载国产大模型，证明中国 AI 产业链的自主可控实力；其二，借 DeepSeek 完善华为云 AI 服务，弥补华为自身盘古大模型在通用领域的短板。

核心需求上，华为希望展示昇腾 AI 芯片在大模型推理上的强兼容与高性能，以及提供安全可控的大模型云服务给政企客户。DeepSeek 开源可部署的技术特点满足

了华为对数据安全的要求（模型可完全运行在华为云本地，不必连接外网），其高性能则可充分利用华为昇腾芯片的算力潜能。

华为还计划将 DeepSeek 模型能力融入其终端和行业解决方案中。例如，在华为云 EI 企业智能、中小企业鲲鹏云等产品线，引入 DeepSeek 用于客户服务、数据分析等场景。

通过接入 DeepSeek，华为打出了国产化与高性能的组合牌：一方面证明“国产芯 + 国产模”完全可以媲美甚至挑战最强海外方案，另一方面也让华为云的 AI 服务能力更上一个台阶，进一步服务其庞大的信息与通信技术生态客户群。

6. 芯片厂商（2025 年 2 月）

值得一提的是，国内大厂接入 DeepSeek 的不仅有云服务和互联网企业，上游芯片公司也纷纷投入支持。在 DeepSeek 发布后的短短一周内，华为昇腾、沐曦科技、壁仞科技、昆仑芯、燧原科技、海光信息、天数智芯等十余家国内 AI 芯片厂商陆续宣布完成对 DeepSeek 模型的适配优化。

这些厂商的出发点在于借 DeepSeek 验证自身芯片的 AI 能力，推动国产硬件在大模型时代的广泛应用。其核心需求是确保自家 GPU / AI 加速卡可以高效运行目前参数量最大、复杂度最高的开源模型之一 DeepSeek，从而向客户证明“我们的芯片也能胜任顶尖 AI 任务”。

DeepSeek 模型的开源和高效特性为它们提供了绝佳试金石。例如，燧原科技（Iluvatar）成功地在其 AI 加速卡上适配了 DeepSeek 全系列模型（包含原生 671B 模型及 1.5B~70B 不同参数量的蒸馏版）；整个过程中燧原的算力得到充分利用，模型推理稳定高效，为后续大规模部署打下基础。

又如，壁仞科技仅用数小时就让其壁砺 GPU 支持了 DeepSeek–R1 的各档蒸馏模型，表现出优秀的兼容性能。这些技术亮点说明 DeepSeek 在设计上具有良好的跨架构适配性，既可运行高端 GPU，也能通过蒸馏压缩在较小算力上流畅运行。这恰恰降低了对进口高端芯片的依赖，为国产芯片提供了用武之地。

国产 AI 芯片公司纷纷拥抱 DeepSeek，一方面可以共享其开源生态红利，获取更多实际应用反馈优化产品；另一方面也借此向市场宣示“我们的芯片 + DeepSeek 方案”可成为替代国外 GPU 的大模型解决方案。

这种上下游协同的景象，正是 DeepSeek 在国内科技产业引发的连锁反应：从云服务到芯片硬件，整个 AI 产业链的玩家都看到了新的机遇与赛道，纷纷参与进来，共建 DeepSeek 的生态朋友圈。