

中国 AI 产业生态与竞争力：冰山与 QZ 模型分析

韩伟

(澳门旅游大学, 中国 澳门, 999078)

摘要: 人工智能技术的规模化商业落地, 标志着全球数字经济进入新的结构性分化时期。本文聚焦中国人工智能产业的竞争格局, 提出“冰山模型”这一分析框架, 将产业生态划分为市场主导层、应用创新层与基础设施层三个维度, 揭示现有研究中普遍被低估的底层资源依赖关系。在此基础上, 本文构建“QZ 三维分析法”(QZ 取 Quantized Zero-to-one 跃迁之意), 从基础力(人才密度与 AGI 信仰)、引力(生态协同能力)及耐力(算力成本效率与商业造血)三个维度评估 AI 企业竞争力, 并设计相应的二级量化指标。冰山模型解决“企业在产业生态中处于哪一位置”的静态结构问题, QZ 三维分析法则回答“企业能否在不同生态位之间跃迁”的动态竞争力问题, 二者共同构成本研究的分析框架。研究以 DeepSeek 系列模型的崛起为案例, 从战略管理视角验证上述框架的解释力。

关键词: 人工智能产业; 冰山模型; QZ 分析法; 数字经济; 战略竞争力

Ecosystem and Competitiveness of China's AI Industry: Iceberg and QZ Model Analysis

Han Wei

(School of Business Administration, Macao University of Tourism, Macao 999078)

Abstract: The large-scale commercialization of artificial intelligence marks a new era of structural differentiation in the global digital economy. This paper proposes the "Iceberg Model" to analyze China's AI industry, categorizing its ecosystem into three layers: market-dominant, application-innovative, and infrastructural. The Iceberg Model addresses the static question of where firms are positioned within the industrial ecosystem, while the "QZ 3D Analysis Method" (QZ standing for Quantized Zero-to-one transition)—built around Foundation, Gravity, and Endurance—addresses the dynamic question of whether and how firms can transition across ecological positions. Using the emergence of DeepSeek as an empirical case, interpreted through a strategic management lens, this paper validates both frameworks and offers practical guidance for corporate digital transformation and regional policy in the Asia-Pacific context.

Keywords: AI Industry; Iceberg Model; QZ Analysis; Digital Economy; Strategic Competitiveness

【作者简介】 韩伟, 男, 澳门旅游大学工商管理博士生, 澳门工商管理博士协会副理事长, 香港未来发展国际集团创始人兼 CEO。主要研究领域: 数字经济、人工智能商业应用、战略管理。

一、引言

2025 年 1 月下旬，一个并不广为人知的中国量化私募机构旗下 AI 团队发布了推理模型 DeepSeek R1，随即在全球资本市场掀起一场始料未及的冲击波，英伟达股价于 1 月 27 日单日蒸发约 5,890 亿美元，创下美股史上单日市值跌幅最大纪录。对于这一事件，多数媒体的解读聚焦于“中美 AI 差距缩小”，但从战略管理的角度看，DeepSeek 现象更深刻的意涵在于：它以一个具体案例，挑战了“算力规模决定竞争壁垒”这一大语言模型兴起以来几乎被奉为公理的行业假设。这一假设的动摇，意味着 AI 产业的竞争格局正在发生结构性重组，而现有的理论工具未必能够有效捕捉这种重组的内在逻辑。

就中国 AI 产业的整体格局而言，2023 年至 2024 年间被频繁以“百模大战”加以描述。然而，这一说法在揭示的同时也遮蔽了一些关键信息：它放大了头部企业之间的同质化竞争，却掩盖了支撑整个产业运行的隐形基础，即那些不在公众视野中出现、却在算力供给、数据资产和算法工具链层面深度介入产业进程的主体。根据中国信息通信研究院的统计^[1]，截至 2024 年底，国内可追踪的大语言模型数量超过 230 个，但具备持续商业落地能力的不超过 20 家。这种极度不对称的生态格局，难以用传统的竞争集中度指标加以充分刻画。当前全球范围内人工智能正从技术验证迈向规模化商业应用（McKinsey Global Institute, 2025）^[2]，这一转折使产业底层的资源依赖结构成为决定竞争走向的关键变量。

现有的 AI 产业竞争力研究在方法论层面存在两个缺口：其一，多数研究沿用 Porter 的五力框架或钻石模型（Porter, 1990）^[3]，难以处理大模型时代“底层模型-应用层-基础设施”之间高度非线性的生态依存关系；其二，以模型能力评测分数或融资额作为竞争力代理指标，忽略了算力获取成本、商业造血能力和生态协同效率等对企业长期生存更具决定性的要素。本文在上述缺口的交叉地带提出两个互补框架：冰山模型解决“企业在哪里”的静态定位问题，QZ 三维分析法解决“企业能否跃迁”的动态评估问题。

二、文献回顾

（一）AI 产业竞争力的理论脉络

早期研究带有明显的宏观技术乐观主义色彩。Brynjolfsson 和 McAfee（2014）^[4]将 AI 定位为继蒸汽机之后的新一代通用目的技术，重点强调其对全要素生产率的提升潜力，而较少深究不同类型企业在这一技术浪潮中的生存差异。Bommasani 等人（2021）^[5]在斯坦福 CRFM 的报告中系统梳理了基础模型的潜力与风险，但关注点仍集中于技术评测，对商业竞争机制分析有限。Liang 等人（2023）^[6]对逾 30 个开源与闭源模型的全面评估表明，模型性能呈现快速均值化趋势，这一发现隐含着重要的竞争学推论：当技术差异化空间压缩，企业竞争的决定因素将从“模型能力本身”转向“能力的成本效率与场景适配”。

与此同时，受 Eisenmann 等人（2006）^[7]平台竞争框架的启发，部分学者将开源策略理解为“吸引开发者—扩大边际效用—强化生态锁定”的正反馈机制，Meta 的 Llama 系列的全球扩散为这一观点提供了有力支持。在中国本土研究方面，戚聿东与肖旭（2020）^[8]从数字经济视角探讨了企业管理变革的底层逻辑；孙早与侯玉琳（2019）^[9]关于工业智能化对就业结构影响的实证研究，从要素市场角度印证了 AI 技术渗透对资源配置的重塑效应，为本文构建 QZ 分析法中的“人才密度”子指标提供了启示。中国信息通信研究院（2023）^[10]的《人工智能大模型技术和应用发展研究报告》在产业全景数据层面具有较高参考价值，但作为行业白皮书，其分析框架缺乏对“竞争边界动态演化”的理论深度。

综合上述文献梳理，现有研究存在三个结构性缺口：缺乏整合多层级生态主体的中观框架；竞争力评估停留于静态指标，缺乏对层间跃迁机制的刻画；对中国 AI 企业的评估多套用美国标准，对“算法替代算力”路径缺乏针对性工具。

（二）本文框架与既有理论的关系

本文两个框架并非对既有理论的替代，而是在特定问题上的延伸与整合。“冰山模型”的理论亲缘是 Schein（1985）^[10]的组织文化冰山隐喻，但本文将分析层次从组织提升至产业，聚焦的是生态系统的结构性可见性偏差而非文化隐性要素；其与 Porter（1990）^[3]五力框架的区别在于，冰山模型揭示的是“哪些行业主体构成了竞争的真实基础”，而非行业竞争态势的整体评估。“QZ 三维分析法”与 Teece 等（1997）^[11]动态能力理论存在深刻的概念共鸣：动态能力的“感知—把握—重构”三元结构，与 QZ 的“基础力—引力—耐力”在功能逻辑上高度对应；QZ 框架的贡献在于，在动态能力的概念空间内设计了能够捕捉 AI 产业特殊竞争逻辑（开源生态、算力成本结构）的可量化子指标，而这是 Teece 框架所未覆盖的维度。与 Eisenmann 等（2006）^[7]平台竞争框架相比，QZ 的引力维度吸收了双边市场逻辑，同时纳入了开源生态这一传统平台理论未充分探讨的协同机制。

在框架定位上需要明确说明：本文的冰山模型与 QZ 三维分析法在当前阶段属于“概念性框架”（Conceptual Framework）。其学术贡献体现在概念界定的清晰度与内在逻辑的自治性，而非即时的量化可复现性。所设计的二级指标（TDI、ACI、OCI、AEP、VCD、CER、BVI）旨在为后续实证研究提供操作化起点，指标权重校准与测量效度的系统验证有待后续研究通过企业面板数据推进。

三、中国 AI 产业生态结构：冰山模型

要理解中国 AI 产业的竞争格局，一个不得不正视的认知误区是：公众和资本市场所能直接感知的那一部分，正是整个生态系统中体量最轻、位置最高的一层。这种“可见性偏差”的存在，使得无论是投资决策还是政策设计，都倾向于高估头部企业的独立性，而低估其对隐形基础设施的深度依赖。

“冰山效应”最初来自冰川学的物理观察，后被 Schein（1985）^[10]引入组织文化研究，用以描述企业文化中可观测要素与深层隐性结构之间的巨大比例差异。本文借用这一隐喻，将分析对象从单一组织延

伸至整个产业，提出“中国 AI 产业冰山模型”，将其划分为三个功能不同、深度依存的结构层级。

本文选择“冰山模型”作为解构中国 AI 产业生态的中观分析工具，而非沿用传统的迈克尔·波特五力框架或行业集中度指标，其核心逻辑根植于中国 AI 产业在特定历史技术窗口期内所呈现的三重异质性特征。

这一选择最直接的方法论价值，在于矫正中国 AI 产业长期存在的“可见性偏差”与舆论泡沫，而这正是聚焦市场可见竞争主体与份额分布的五力模型及集中度指标难以摆脱的视觉牵引。自 2023 年“百模大战”爆发以来，媒体与资本市场的注意力多聚焦于市场主导层的参数竞赛与前端流量，而这种显性的繁荣掩盖了中国 AI 产业极度不对称的生态依赖。冰山模型通过“水面上/下”的结构划分，强制将研究视线引向那些隐没于公众视野之外的算法工具链、国产芯片适配生态以及行业私有数据集，从而能够更真实地评估产业的整体韧性。

更深层的契合性，在于该模型能够捕捉外部地缘政治约束下中国 AI 产业“非对称依赖”的底层现实，而这正是聚焦同层横向竞争关系的五力模型先天难以触及的纵向命题。在前沿硬件获取遭遇壁垒的背景下，中国 AI 企业无法单纯复制西方依赖充沛算力堆叠的平铺式路径；DeepSeek 等案例表明，中国 AI 的战略突围高度依赖于基石层中极端算法优化对硬件性能缺陷的非对称补偿。冰山模型提供的“底部支撑上层”纵向视角，能够刻画这种知识底蕴与工程红利对上层商业创新的非对称拓扑关系。

与此同时，该模型还能够容纳中国互联网大厂独特的“跨层纵向一体化”生态，而这正是依赖清晰边界划分的集中度指标的盲区所在。区别于西方相对清晰的平台分工，中国 AI 生态中的头部平台（如阿里云、腾讯云等）往往同时横跨基石层（算力租赁）、主导层（通用大模型）与应用层。冰山模型通过对动态层级的判定与跨层主体的界定，避免了传统分业分析工具对生态系统完整性的割裂，为后续 QZ 三维分析法中“层间跃迁”的动态演进提供了稳固的静态坐标。

（一）冰山之巅：市场主导层

这一层级的主体，是拥有最大品牌曝光度、资本吸引力和模型研发能力的头部企业群体。主要包含两类：大型互联网平台旗下的 AI 战略部门（百度文心一言、阿里通义千问、腾讯混元、字节豆包）与获得大规模外部融资的大模型初创企业（月之暗面、智谱 AI、商汤日日新）。

从融资数据看，这一层级在 2023 至 2024 年间吸引了市场的绝对注意力：月之暗面于 2024 年 2 月完成的 A+ 轮融资金额超 10 亿美元，投后估值达 25 亿美元，并于同年 8 月完成超 3 亿美元 B 轮融资，估值进一步升至 33 亿美元；智谱 AI 于 2024 年 12 月完成单轮 30 亿元人民币融资，投后估值突破 200 亿元^[12]。模型性能层面，头部企业在 Chatbot Arena 等国际基准评测上与 OpenAI、Google 的差距在 2024 年下半年已大幅收窄。然而，商业变现路径的相对单一，主要集中于 API 调用、企业私有化部署及云计算绑定，这始终是这一层级企业的核心生存压力所在。

（二）水面过渡层：应用创新层

应用创新层是目前中国 AI 生态中企业数量最多、竞争逻辑最为多元的层级，其共同特征是：不追求在通用底层模型上与巨头正面竞争，而是以大模型 API 为底座，向特定垂直场景深度切入，以场景数据积累为核心护城河。

金融科技领域是落地程度相对较高的代表性赛道。东方财富、同花顺等平台已将 AI 功能深度嵌入核心业务流程。以同花顺为例，其月活跃用户规模在 2024 年下半年持续稳居证券类应用首位（易观千帆数据）^[13]，高粘性、高频次的用户基础为 AI 功能的规模化渗透提供了天然入口。医疗健康领域，科大讯飞医疗的 AI 辅助诊断系统截至 2024 年底已覆盖全国超 7 万家基层医疗机构，并接入 500 余家等级医院^[14]，借助持续的场景积累在标注数据层面形成了较高的进入壁垒。应用创新层中另一个值得关注的子群体，是以技术时间差为核心竞争资产的“出海先锋”。受中美科技摩擦影响，部分中国 AI 应用企业选择以东南亚、中东及拉丁美洲为起点，在算法生态尚不成熟的市场建立先发优势。

（三）冰山基石层：隐形基础设施

基石层是整个中国 AI 产业得以持续运转的物质基础与知识源头，却也是最难被外部观察者所充分感知的部分。这一层级涵盖高校与科研机构、算力基础设施提供者、高质量数据生态建设者三类主体，三者的关系并非简单的供应链上下游，而是相互咬合的功能性依存。根据 IDC《中国人工智能算力发展评估报告（2025）》^[15]，截至 2024 年底，中国智能算力总规模已达 725.3 EFLOPS，同比增长 74.1%，以华为升腾 910B 为代表的国产算力芯片在适配效率上持续提升，支撑了在出口管制背景下国内大模型预训练的基础需求。

理解基石层的战略意义，需要引入“供应链韧性”这一视角。DeepSeek 能够在 H800 GPU（而非受出口管制的 H100）条件下完成高质量模型训练，其技术前提是对芯片底层指令集和显存利用率的极端优化^[16]，而这一能力的形成，依赖于基石层长期积累的算法工程底蕴。换言之，表面上看是市场主导层的企业完成了技术创新，但支撑这一创新的知识基础，根植于整个产业生态系统最不显眼的部分。

（四）层级判定标准：初步阈值建议

为增强不同研究者之间的可复现性，本文为三个层级分别提出初步的判定阈值。市场主导层的判定采用多条件组合：在累计融资额超过 10 亿元人民币、旗下通用大模型于 Chatbot Arena 或等效评测中进入全球前 20 名、月均 API 调用量超过 10 亿次这三项指标中，满足任意两项即可归入该层；考虑到大型平台旗下的 AI 部门通常不单独披露融资数据，对其判定可主要依据后两项指标。应用创新层的核心识别特征，是以垂直场景应用为主要商业模式：这类企业在至少一个细分领域拥有逾百万月活跃用户或逾亿元年度 B2B 合同额，且技术路线主要依赖外部模型 API 调用，而非自主预训练。基石层的核心识别特征，则是其主要收入来源为算力租赁、数据服务或底层工具链，而非面向终端用户的 AI 产品。需要特别说明的是，部分企业同时在多个层级布局，宜视为“跨层主体”，而非强行归入单一类别。例如，

阿里云既提供算力租赁（基石层），又推出通用大模型（主导层）；科大讯飞则一方面以自主预训练的星火大模型布局主导层，另一方面通过讯飞医疗等产品深耕应用层场景。

四、企业核心竞争力评价：QZ 三维分析法

冰山模型解决的是一个静态的定位问题：企业在中国 AI 产业的三层生态结构中处于何种位置，其竞争逻辑与资源依赖是什么。但仅仅知道“企业在哪里”，还不足以回答“企业能否在不同生态位之间成功跃迁”这一更具实践张力的问题。DeepSeek 从基石层向市场主导层的成功跃迁，正是对这一问题的有力提示。因此，本文在冰山模型的基础上进一步构建“QZ 三维分析法”，作为评估企业动态竞争力的补充工具。

QZ 三维分析法的设计初衷，是在可操作性与解释力之间寻求平衡。“QZ”取“Quantized Zero-to-one 跃迁”之意，暗喻 AI 企业从量变积累到质变跃升的竞争动力学过程。本文识别出三个对 AI 企业长期竞争力具有普遍解释力的核心维度，并在每个维度下设计具体的二级量化指标，使得跨层级、跨类型的企业横向比较成为可能。

（一）基础维度（Foundation）：人才密度与 AGI 信仰

基础维度含两个互补子指标。“人才密度指数”（TDI）关注顶尖研究者的集聚效应而非员工总数，代理指标建议采用：过去三年在顶级 AI 学术会议（NeurIPS、ICML、ACL、ICLR）中一作或通讯作者论文数、高被引研究者留存率及核心技术岗薪酬溢价率。DeepSeek 以不超过 50 人的核心算法团队在两年内发表逾 30 篇顶会论文，是这一“少而精”战略的典型体现。“AGI 信仰系数”（ACI）评估企业能否在商业压力下维持对长期技术路线的战略定力，代理指标包括研发投入占收入比重及公开技术承诺与资源配置的一致性。ACI 的缺失，往往是企业向短期“套壳”策略滑落的前兆。

（二）引力维度（Gravity）：生态协同能力

引力维度描述企业在技术生态中的向心力，其本质是“以开放换增长”的平台竞争逻辑：让渡部分技术控制权换取开发者社区的规模效应，以低边际成本获取场景数据形成数据飞轮。二级指标分三层：“开源贡献度”（OCI）通过 GitHub 及 HuggingFace 的可量化数据评估，涵盖 Star 增长速率、外部贡献者活跃度及模型下载量；“API 生态渗透率”（AEP）衡量模型能力向下游场景渗透的广度；“垂直场景覆盖密度”（VCD）评估企业在医疗、金融、教育、政务等落地场景的深度参与。VCD 刻意区分“全景布局”与“场景深耕”，后者对构筑数据壁垒更具决定性意义。

（三）耐力维度（Endurance）：算力成本效率与商业造血

耐力维度解决企业的根本性生存问题：在长期技术投入的同时是否维持足够的财务生命力。“算力效率比”（CER）的核心逻辑是：在达到相同性能基准的条件下，算力成本越低，现金流优势越显著。

以 DeepSeek V3.2 为基准，其 API 输入定价约为\$0.28/百万 tokens，相较 GPT-5.4（\$1.25/M）低 78%，相较 Claude Opus 4.6（\$15/M）低逾 98%；DSA 架构在长文本场景下进一步将推理成本压缩逾 50%^{[16][17]}。“商业造血指数”（BVI）重点考察 ToB 私有化部署的客单价与续费率、开放 API 的计费转化率，以及是否已建立独立于外部融资的正向现金流循环。BVI 偏低的企业在融资环境收紧时生存韧性将受严峻考验。

（四）QZ 框架的初步应用：代表性企业横向评分

为验证 QZ 框架的区分效度，本文依据公开信息对智谱 AI、月之暗面、MiniMax 及 DeepSeek 四家企业进行初步评分（五星制）。此为定性估算，旨在展示框架的区分逻辑而非精确量化结论。

表 1 代表性中国 AI 企业 QZ 三维分析法初步评分（基于公开信息，2025 年）

企业	基础维度（TDI+ACI）	引力维度（OCI+AEP）	耐力维度（CER+BVI）	综合评级
DeepSeek	★★★★★	★★★★★	★★★★☆	旗舰型
智谱 AI	★★★★☆	★★★★☆	★★★☆☆	均衡型
月之暗面	★★★★☆	★★★☆☆	★★☆☆☆	基础依赖型
MiniMax	★★★☆☆	★★★★☆	★★★☆☆	场景深耕型

注：★★★★★= 极强；★★★★☆= 强；★★★☆☆= 中等；★★☆☆☆= 偏弱。数据截至 2025 年底。

从评分分布看，QZ 框架呈现出较好的区分效度。DeepSeek 在基础与引力维度均达最高评级，验证了其“算法精英 + 开源生态”双轮驱动模式；智谱 AI 三维评分相对均衡，整体布局稳健但尚未形成压倒性优势；月之暗面的耐力维度偏弱，折射出高融资依赖与商业造血不足的结构脆弱；MiniMax 通过多模态场景深度覆盖在引力维度表现突出，但基础维度的研究积累相对有限。这一格局与行业观察基本吻合，初步支持了 QZ 框架作为区分性分析工具的有效性。

五、案例研究：DeepSeek 竞争能力的持续演进与 QZ 框架验证

在展开案例分析之前，有必要就方法选择作出说明。本文之所以采用单案例研究（single-case study）而非大样本统计检验，根植于研究问题的性质与研究对象的现实条件。其一，本文的核心任务是检验冰山模型与 QZ 三维分析法这两个新建构框架的解释力，属于 Yin（2018）^[18]所界定的“理论检验型”案例研究，即当研究目标是回答“如何（how）”与“为何（why）”的过程性问题，而非“多少（how much）”的频率性问题时，案例研究相较于变量统计更具适配性。其二，DeepSeek 构成了一个 Eisenhardt（1989）^[19]意义上的“极端案例”（extreme case）：它在硬件受限的约束条件下完成了从基石层向市场主导层的逆向跃迁，这一现象在现有样本中尚属罕见，其信息密度远高于均值化的普通样本，能够充分暴露框架所刻画的层间跃迁机制。其三，AI 产业正处于剧烈的范式转换期，可供统计的层间跃迁事件样本量过小且高度异质，尚不具备大样本回归的统计前提；在此阶段，深度案例研究是建构与验证中观理论更稳健的方法路径。此外，本文亦引入“未遂跃迁”的失败案例群（见本节第五部分）作为反向对照，以缓解单案例研究的外部效度局限。

此外，鉴于人工智能模型的迭代速率远超传统行业，有必要明确本文的数据截断时点（data cutoff）。本文所涉及的全部模型版本、性能评测指标与 API 定价数据，均以 2026 年第一季度末（2026 年 3 月 31 日）为统一截断时点；该时点之后发布的后续模型不纳入本文的经验分析范围。作出这一界定，并非忽视技术演进，而是出于实证研究的内在要求：唯有冻结一个明确的时间断面，全文数据方能在同一时点上保持可比性与可复现性。需要强调的是，本文的核心论证建立在已被市场充分验证的既成事件之上（DeepSeek-R1 与 DeepSeek-V3.2 的发布及其引发的产业链反应），而非依赖于某一特定时点的最新模型；因此，截断时点之后的模型演进，并不影响本文核心结论的成立，反而可作为对本文分析框架解释力的持续检验（参见第六节的相关讨论）。

（一）事件背景：从 R1 冲击到 V3.2 确立

2025 年 1 月 20 日，DeepSeek 发布推理模型 DeepSeek-R1 并同步开源模型权重，成为当年 AI 产业最具冲击力的事件节点。在能力评测层面，DeepSeek-R1 在 AIME 2024 测试中得分 72.6%，与 OpenAI o1 正式版的 74.4% 相当；在 Codeforces 的代码能力评测中，其 ELO 评分达到 1,820 分，进入全球顶尖程序员前 0.3% 分位段^[20]。英伟达股价于 1 月 27 日单日蒸发约 5,890 亿美元，创下美股史上单日市值跌幅最大纪录。

2025 年 12 月 1 日，DeepSeek 正式发布 DeepSeek-V3.2（及其强推理变体 V3.2-Speciale），以完全开源的方式对外发布 685 亿参数 MoE 架构模型。V3.2 在 2025 年国际数学奥林匹克（IMO）和国际信息学奥林匹克（IOI）中取得金牌成绩，性能指标与同期 GPT-5 相当^[16]，而 API 输入定价仅为约 0.28 美元 / 百万 tokens，相较于 GPT-5.4 的约 1.25 美元 / 百万 tokens^[17]，差距超过四倍。这一结果表明，R1 所开创的“算法效率替代算力规模”路径，已从一次性突破演变为可复制的技术方法论。

这一市场反应需要被准确解读：资本市场的震动，并非来自对 DeepSeek 本身商业前景的直接估值，而是来自对“高端 GPU 长期稀缺性溢价”的重新定价。V3.2 的发布进一步印证了这一判断：以 DeepSeek Sparse Attention（DSA）为代表的新一代算法架构创新，在长文本推理场景中将计算成本较前代压缩逾 50%，同时维持了性能基准的持平甚至小幅提升。

为直观呈现截至 2026 年第一季度“算力效率比”指标的跨模型差异，下表对 DeepSeek V3.2 与同期主要前沿竞品进行横向对比：

表 2 主要前沿大模型核心竞争力指标横向对比（2026 年 Q1）

指 标	DeepSeek V3.2	GPT-5.4	Claude Opus 4.6
发布时间	2025 年 12 月	2026 年 3 月	2026 年 2 月
训练成本（美元）	约 600 万（估）	不公开（估逾亿）	不公开
API 输入定价（\$/ 百万 token）	约 0.28	约 1.25	约 15.0
AIME 2025 数学推理	89.3%（V3.2-Exp）	94.6%（GPT-5）	未公开
SWE-bench 代码能力	具备竞争力（待验证）	74.9%+	80.8%
算力效率比（CER）	★★★★★	★★★	★★
开源策略	完全开源（MIT）	闭源	闭源

资料来源：DeepSeek 技术报告（2025）、OpenAI GPT-5.4 发布公告（2026-03-05）、Anthropic Claude 4.6 基准评测（2026-02-05）及公开 API 定价整理。注：本文数据统一截至 2026 年第一季度末（2026 年 3 月 31 日），表中所列 GPT-5.4、Claude Opus 4.6 均为截断时点前已正式发布之版本，并非预测数据；该时点之后发布的后续模型不纳入本文经验分析范围（详见第五节说明）。

（二）冰山模型定位：从基石层到市场主导层的跃迁

DeepSeek 的崛起路径，是冰山模型“层间跃迁”逻辑的典型实例。DeepSeek 隶属于量化私募机构幻方科技旗下，后者的核心竞争力根植于高频交易中的数学建模和算法优化，这正是本文所定义的“基石层”能力范畴：深度的算法工程底蕴和高密度的数理人才储备，而非亿级参数规模的大模型研发组织。DeepSeek 正是以这种“基石层”的能力禀赋，通过算法效率的突破，完成了向市场主导层的成功跃迁。

这一跃迁路径的战略意义在于，它提供了一种不依赖算力资源积累的竞争替代路径，这与克里斯坦森（Christensen, 1997）^[21]的“破坏性创新”（Disruptive Innovation）逻辑高度契合：在技术范式发生根本性转换的窗口期，以低成本、高效率的新维度重新定义竞争基础，往往比沿既有技术轨道延续性投入更具颠覆力量。DeepSeek 以算法工程效率为突破口，实质上完成了一次典型的破坏性进入，其竞争优势并非建立在现有游戏规则上的超越，而是通过改写规则本身实现的跃迁。

（三）以 QZ 三维分析法解构 DeepSeek 的竞争优势

从基础维度审视，DeepSeek 的人才密度指数（TDI）在国内 AI 企业中处于顶尖水平。其核心算法团队不超过 50 人，但在 2023 至 2024 年间以一作或通讯作者身份发表顶会论文逾 30 篇。尤为关键的是，幻方科技的量化投资背景，塑造了这支团队对算法精度和计算成本的极端敏感性。这种对“效率”的偏执，在大厂以工程量和数据规模取胜的研发文化中殊为罕见，构成了一种难以被快速复制的组织能力。

从耐力维度看，DeepSeek 从 R1 到 V3.2 的持续选择，是以“算法创新降低资源消耗”代替“扩大资源投入维持竞争力”，在技术封锁背景下贯彻执行成本领先战略。就具体成效而言，V3.2 的 API 定价（\$0.28/百万 tokens）相较于 GPT-5.4（\$1.25 / 百万 tokens）低约 78%，相较于 Claude Opus 4.6（\$15 / 百万 tokens）更构成数量级差距。DSA 架构创新将长文本场景的推理成本压缩逾 50%，进一步强化了这一耐力优势的可持续性。

从引力维度看，DeepSeek 从 R1 到 V3.2 一贯之地执行开源策略，带有明确的主动设计逻辑而非被动选择。完全开源（包括模型架构、训练代码及技术报告，MIT 协议）的决策，可被理解作为一种“以知识产权换生态规模”的平台策略。DeepSeek R1 技术报告发布后两周内，已出现逾 50 个基于其蒸馏技术的垂直领域衍生模型；而 V3.2 的开源释出，进一步在 HuggingFace 上引发大规模的社区复现与二次开发，其模型仓库的 Fork 数量在发布后 30 日内突破 2 万次^{[20][16]}，DeepSeek 正从一个中国 AI 企业，逐步演化为具有全球影响力的开源 AI 生态节点。

（四）DeepSeek 跃迁路径的边界条件

分析 DeepSeek 成功的同时，指出其可复制性的边界，对本研究框架的完整性同样重要。DeepSeek 的算力效率优势，深度依赖母机构在量化金融领域积累的数学优化文化，这一类型的组织基因，难以通过快速招聘或短期培训加以复制，更接近于 Barney（1991）^[22]所定义的异质性稀缺资源。此外，DeepSeek 在商业造血指数（BVI）这一子维度上目前的表现尚待市场的进一步检验。这一观察反过来印证了 QZ 三维分析法的设计初衷：三个维度需要同时评估、相互印证，任何一个维度的突出表现都不足以保证企业的长期竞争优势。

（五）失败案例对照：“未遂跃迁”的 QZ 解读

理论框架的解释力不仅体现于对成功案例的描述，更体现于对失败的预测能力。2023 至 2024 年间，有多家中国 AI 创业企业尝试从应用创新层向市场主导层跃迁，最终因资源耗尽而退出。以 QZ 框架分析，这些企业的共同特征是：高管层背景多来自互联网运营岗位，顶会论文积累严重不足（TDI 偏弱）；研发投入占比持续被商业扩张稀释（ACI 缺失）；通用大模型的预训练成本往往数倍于其单轮融资规模（CER 极低）。与 DeepSeek 从基石层成功跃迁的路径对比，这一“未遂跃迁”的模式进一步支持了本文的核心论断：决定层间跃迁成败的关键是基础维度的算法能力积累，而非耐力维度的短期资金规模。

六、结论与管理启示

本文所提出的冰山模型与 QZ 三维分析法，试图从不同角度回应同一个核心问题：在技术迭代速率远超传统行业的 AI 产业中，企业竞争力的来源究竟是什么，应该如何被测量和比较？冰山模型澄清了一个认知误区：在中国 AI 产业中，可见度最高的企业未必是决定产业走向最关键的节点，基础设施层的芯片生态、数据资产与算法人才储备，以更缓慢但更深刻的方式塑造着上层竞争的边界条件。QZ 三维分析法则提供了一套相对统一的评估语言，使得跨层级、跨类型的企业动态比较成为可能。

对企业管理者而言，上述框架最直接的启示是“生态位自知”的重要性。在制定 AI 转型战略之前，企业需要诚实评估自身在冰山模型中所处的位置：试图在市场主导层与巨头正面竞争、在应用创新层寻找场景纵深，还是深耕基石层的特定节点。这三种选择背后所需的资源结构和能力组合，有着本质的差异，而混淆这三者的企业往往在三个方向上都没能形成真正的壁垒。

对政策制定者而言，冰山模型所揭示的基础设施依赖关系，意味着产业扶持政策的着力点不应只聚焦于“明星大模型的参数竞赛”。DeepSeek 的案例表明，在硬件获取存在外部约束的现实背景下，算法效率优化的政策引导与资源支持，在成本效益上可能优于算力规模的单纯扩张。具体而言，建议在“新一代人工智能发展规划”的后续实施中，将算法效率优化研究列为专项资金支持方向，并推动高校量化工程、计算机科学与管理科学的跨学科人才培养机制。

值得补充的是，本文截稿后的产业动态为上述框架提供了一次前瞻性验证。2026 年 4 月，DeepSeek 发布新一代 V4 系列模型（V4-Pro 与 V4-Flash）^[23]，延续了完全开源（MIT 协议）与效率优先的一贯路线：其混合稀疏注意力架构在百万 token 长上下文场景下，将单 token 推理算力压缩至前代 V3.2 的约四分之一，同时以约为同期闭源前沿模型十分之一量级的 API 价格，维持了具有竞争力的性能表现。这一演进虽不属于本文截断时点（2026 年第一季度末）内的经验分析对象，但其内在逻辑，即以算法与架构效率的持续突破替代算力规模的单纯堆叠，恰是本文所识别的“算法效率替代算力规模”路径的延续与强化。换言之，V4 的发布并未改变本文的核心论断，反而从框架外部印证了 QZ 三维分析法所揭示的耐力维度（算力成本效率）作为中国 AI 企业可持续竞争优势来源的稳健性。这也从侧面说明，一个具有解释力的中观分析框架，其价值不仅在于解释既有现象，更在于对产业演进方向的前瞻预测能力。

本研究的局限性有必要坦诚说明。冰山模型中各层级的划分标准目前仍依赖研究者的定性判断，尚未发展出严格的量化界定方法；QZ 三维分析法的二级指标权重，也有待通过更大样本的企业面板数据进行实证校准。后续研究可在以下方向进一步推进：基于纵向数据的 AI 企业动态竞争力追踪，开源策略对市场份额长期影响的因果识别，以及冰山模型在东南亚、中东等新兴 AI 市场的适用性验证。

【参考文献】

- [1] 中国信息通信研究院。人工智能大模型技术和应用发展研究报告（2023 年）[R]。北京：中国信通院，2023。
- [2] McKinsey Global Institute. The State of AI in 2025: AI's Breakout Year in Business[R]. McKinsey & Company, 2025.
- [3] Porter M E. The Competitive Advantage of Nations[M]. New York: Free Press, 1990.
- [4] Brynjolfsson E, McAfee A. The Second Machine Age[M]. New York: W.W. Norton, 2014.
- [5] Bommasani R, Hudson D A, Aditi E, et al. On the Opportunities and Risks of Foundation Models[R]. Stanford CRFM, 2021.
- [6] Liang P, Bommasani R, Lee T, et al. Holistic Evaluation of Language Models[J]. Transactions on Machine Learning Research, 2023.
- [7] Eisenmann T, Parker G, Van Alstyne M. Strategies for Two-sided Markets[J]. Harvard Business Review, 2006, 84(10): 92-101.
- [8] 戚聿东，肖旭。数字经济时代的企业管理变革 [J]. 管理世界，2020, 36 (6): 135-152.
- [9] 孙早，侯玉琳。工业智能化如何重塑劳动力就业结构 [J]. 中国工业经济，2019 (5): 61-79.
- [10] Schein E H. Organizational Culture and Leadership[M]. San Francisco: Jossey-Bass, 1985.
- [11] Teece D J, Pisano G, Shuen A. Dynamic Capabilities and Strategic Management[J]. Strategic Management Journal, 1997, 18(7): 509-533.
- [12] IT 桔子。IT 桔子创业投资数据库：中国人工智能领域投融资交易数据[DB/OL]. [2026-06-17]. <https://www.itjuzi.com>.
- [13] 易观分析。易观千帆：2024 年证券应用类 APP 活跃度数据[EB/OL]. (2025-01-19)[2026-06-17]. <https://www.analysys.cn>.
- [14] 讯飞医疗科技股份有限公司。截至 2024 年 12 月 31 日止年度之全年业绩公告[R/OL]. (2025-03-26)[2026-06-17]. 香港：香港联合交易所，2025.
- [15] IDC, 浪潮信息. 2025 年中国人工智能算力发展评估报告 [R]. 北京: IDC 中国，2025-02-13.
- [16] DeepSeek-AI. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models[R]. Technical Report, 2025.
- [17] OpenAI. Introducing GPT-5.4[EB/OL]. <https://openai.com/index/introducing-gpt-5-4/>, 2026-03-05.
- [18] Yin R K. Case Study Research and Applications: Design and Methods[M]. 6th ed. Thousand Oaks, CA: SAGE Publications, 2018.
- [19] Eisenhardt K M. Building Theories from Case Study Research[J]. Academy of Management Review, 1989, 14(4): 532-550.
- [20] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[R]. Technical Report, 2025.
- [21] Christensen C M. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail[M]. Boston: Harvard Business School Press, 1997.
- [22] Barney J B. Firm Resources and Sustained Competitive Advantage[J]. Journal of Management, 1991, 17(1): 99-120.
- [23] DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence[R]. Technical Report, 2026.