



# 国产大模型的效率 革命

DeepSeek-V4 引领 AI 工业化落地新纪元

北京研精毕智信息咨询有限公司

# 目录

1

技术解构：三大突破

2

商业化分析：性价比为王

3

行业竞争新格局

4

未来展望：起点而非终点



## 技术解构：三大突破

DeepSeek-V4 以“可控性 + 低成本 + 场景适配”为核心，通过 MoE 架构优化、上下文窗口扩容及全链路国产化适配，实现了技术叙事与商业价值的双向突破。

# 双版本 MoE 架构

打破“参数与效率”的矛盾，实现弹性算力调度

1

## 动态路由机制

采用“基础版 8 专家 + 旗舰版 16 专家”设计，针对不同场景智能调度算力资源。

2

## 弹性算力匹配

日常任务仅激活 20% 参数，复杂任务全量调用，精准匹配企业“弹性算力需求”。

3

## 性能与成本双赢

专家切换延迟降低 37%，长文档摘要准确率提升 29%，算力消耗减少 42%。

4

## 顶尖编程能力

V4-Pro-Max 在 LiveCodeBench 斩获 93.5 分，Codeforces 评级 3206，跻身人类选手前 23 位。

# 百万 Token 上下文

## 从“文本理解”到“场景建模”的飞跃

1

### 技术突破

通过“分段注意力机制 + 动态显存管理”，首次实现 100 万 Token 无衰减处理。

2

### 处理能力

相当于一次性理解 20 本《三体》，长距离推理准确率保持在 85% 以上。

3

### 场景落地

在法律卷宗分析、工业设计文档拆解、金融研报整合等场景效率提升 10 倍以上。

4

### 实用化局限

超过 128K Token 后检索准确率开始下降，100 万 Token 时仅为 66%，仍需优化。



# 全链路国产化适配



## 破解“卡脖子”痛点，重塑产业格局

01

### 广泛兼容性

全面兼容华为昇腾 910B、海光 DCU、寒武纪思元 370 等国产芯片及麒麟、统信等系统。

02

### 核心选型考量

全链路适配能力正成为政企客户选型的核心因素，市场需求增速达 189%。

03

### 卓越实测表现

在昇腾 910B 芯片上，推理性能达同配置英伟达 A100 的 89%，训练效率达 83%。

04

### 深远产业影响

推动国产算力产业链爆发，重塑 AI 产业地缘竞争格局。





# 商业化分析：性价比为王

DeepSeek-V4 以“分层定价 + 生态布局”为核心，直击 B 端“成本控制”与“场景落地”的核心需求，将极致性价比变为国产大模型突围的关键武器。

# 分层定价策略



## 打破“高价壁垒”，覆盖全场景客户

01

### 首要决策因素

市场调研显示，“成本控制”已超越“技术性能”，成为企业选型首要考量。

02

### 极致性价比

V4-Flash 版本输入仅 0.14 美元 / 百万 Token，输出 0.28 美元，较 GPT-4 和 Claude Opus 便宜 7 倍以上。

03

### 灵活套餐设计

针对中小企业按调用次数计费，无最低消费；中大型企业采用年费专属部署套餐，价格降低 50% 以上。

04

### 开发者友好

开发者与科研机构可享受每月 1000 万 Token 免费额度，大幅降低使用门槛。



# 生态布局矩阵

## 从“单一模型”到“解决方案矩阵”的升级



1

### 垂直场景深耕

针对金融、法律、医疗、工业推出“模型 + 工具 + 数据”一体化方案，客户留存率高 47%。

2

### 标杆客户签约

已签约 20 余家标杆客户，金融“研报生成 + 风险预警”、法律“合同审查 + 案例检索”系统表现亮眼。

3

### 开放开发者生态

开放模型微调平台 DeepSeek FineTune，提供低代码工具链，支持企业用最低 100 条样本定制专属模型。

4

### 生态地位巩固

超 3 万名开发者入驻，累计创建 2.3 万个定制化模型，在中国企业级大模型三强格局中稳居第三。





# 行业竞争新格局

DeepSeek-V4 的出现，标志着国产大模型从“跟跑”向“并跑”的关键跨越，推动行业从“参数竞赛”全面迈入“效率优化”与“差异化竞争”的新阶段。

# 核心对手对比



## 技术与商业双维度下的差异化竞争

01

### DeepSeek-V4

核心优势为长上下文、全链路国产化与高性价比，短板是垂直场景数据积累尚浅，市场份额 18.4%。

02

### 文心一言 4.0

依托百度生态，多模态能力突出，但长文本处理仅支持 32 万 Token，难以满足大规模文档需求。

03

### 通义千问 3.0

阿里生态加持，行业解决方案成熟，以 32.1% 的市场份额领跑，但国产化算力适配滞后。

04

### GPT-4 Turbo

多模态与生态完善，但价格高且无国产化适配，难以满足国内政企合规要求。

# 未来趋势判断

## 三大方向成竞争焦点，效率为王

### 01 技术层面

从“参数竞赛”转向“效率优化”，长上下文、低延迟、国产化适配成为核心比拼点。

### 02 商业层面

“模型免费 + 服务收费”模式普及，垂直场景解决方案的落地能力决定市场份额。

### 03 生态层面

“算力 + 模型 + 工具 + 数据”闭环生态成为核心壁垒，仅靠单一环节优势难以长期立足。



# 未来展望：起点而非终点

DeepSeek-V4 的发布，是国产大模型聚焦企业落地痛点的“实用型产品”起点，它开启了 AI 工业化落地的“拐点时刻”，预示着一个以“价值创造”为核心的万亿级赛道正加速成型。

# 效率革命的拐点



## 聚焦痛点，开启 AI 工业化落地新纪元

01

### 市场规模预判

中国 AI 大模型市场规模预计 2026 年将突破 700 亿元，复合增长率高达 116%，行业迎来黄金发展期。

02

### 关键跨越标志

V4 的出现，标志着国产大模型从“跟跑”向“并跑”的关键跨越，是起点而非终点。

03

### 核心价值回归

未来大模型的竞争终将回归“价值创造”的本质——谁能更好地解决实际问题，谁就能站稳脚跟。



# 下一站：超越之路



## 持续突破，迈向“价值创造”的本质

1

### 核心突破方向

要实现“超越”，国产模型还需在数据质量、生态完善、盈利模式上持续突破。

2

### 企业的绝佳机会

V4 的低价策略与国产化适配，为企业发展提供了“低成本试错”的绝佳机会。

3

### 行业的加速器

这场“效率革命”将加速 AI 在各行各业的规模化落地，重塑生产与服务模式。



## 分析师声明

负责本研究报告的分析师在本报告中所采用的数据均来自合规渠道，报告的观点、逻辑和论据均为分析师本人研究成果，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

## 公司声明

本报告的著作权归北京精毕智信息咨询有限公司（简称为“研精毕智”）所有。本报告是研精毕智研究与统计成果，所载的观点、结论

和建议仅代表行业基本状况，仅为市场及客户提供基本参考。  
本报告调研方法主要是桌面研究、行业访谈等，结合公司内部逻辑算法，通过定量和定性分析分析，客观阐述行业的现状，科学预测

行业未来的发展趋势。  
我们力求报告内容客观、公正，但受到调研方法及调查资料收集范围的局限，本报告所述的观点、数据并不一定完全准确。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式篡改、复制和发布。如引用、转载需注明出处，且不得

对本报告进行有悖原意的引用和修改。  
本研究报告仅供北京研精毕智信息咨询有限公司客户和经本公司授权机构的客户使用，未经授权私自刊载的机构以及其阅读和使用

应慎重使用报告，本公司不承担由此所产生的相关风险和责任。