

## 行业研究

## 互联网大厂如何受益于 DeepSeek-R1 “破圈”？

## ——互联网行业专题研究

## 要点

**DeepSeek-R1 “破圈” 拆解：**1、业界领先的强大性能。1) R1 在多个基准测试中的表现超越当下全球 AI 行业领先的推理模型 OpenAI-o1。2) 在开发人员和使用中收获高评价，在 Chatbot Arena 榜单中居前列，超过 OpenAI-o1。3) 英文日常问答、物理测试等实际用户体验不亚于 o1 系列。2、**多项算法和工程上的实质性突破。**首个验证后训练时使用强化学习让千亿参数的模型获得推理能力的研究，切实解决行业难题。经历多代模型，R1 实现在 GRPO 算法、MoE 架构、MLA 机制、FP8 精度、MTP 方法等多方位突破。3、**全面开源并推出免费 C 端产品。**DeepSeek App 成为大部分用户首次体验的优质 AI 推理模型，在几乎没有广告投放情况下 7 天用户增长 1 亿。

**互联网大厂大模型进展梳理：**1、**阿里巴巴：**1) 旗舰模型 Qwen2.5-Max：指令模型、基座模型的指标对比中赶超业界领先模型，代码编写等能力、应用体验提升。2) 实验性推理模型：数学和编程等领域取得进步，期待新模型赋能、DeepSeek-R1 技术启示，正式版带来突破。3) 开源：提供全尺寸开源模型，性能和开发者参与度均领先。2、**腾讯：**1) 架构调整：TEG 聚焦技术底座，其他事业群共推产品化，腾讯 AI 产品 25 年整体进展有望相较 24 年更积极。元宝并入 CSIG 或显示出腾讯 AI 战略转变更加重视产品体验。2) 多模态较多进展：业界首个一站式 3D 内容 AI 创作平台。支持游戏等工作流，几何结构更精细，纹理色彩更丰富。3、**百度：**1) 文心 4.0 Turbo 24M6 发布，期待 25 年新版本。2) 两大产品助力全栈服务解决方案：大模型精调和应用开发平台千帆，提供稳定高效算力服务的百舸。4、**快手：**文生视频生成模型可灵始终处于全球业界领先水平，最新基座版本更新后，带来显著画面表现力提升，并获专家评测榜单好评。5、**美团：**联合推出 MobileVLM V2 模型，有望在移动环境占优。6、**哔哩哔哩：**Index 自研模型角色扮演等能力不俗。

**DeepSeek “破圈” 后，对互联网大厂有何价值？**1、**大模型能力提升：**R1 带来的突破有机会持续完善，推出性能更强大的模型。R1 的模式有助于激发现有模型潜力，如对阿里 Qwen2.5 等模型进行微调。DeepSeek 的成功或有望促使各大互联网公司加大对 AI 大模型的战略投入。R1 多项创新性技术突破路径已开源，其或可被复刻至各大互联网公司旗下 AI 模型中，带来模型能力的提升，进一步加强对内赋能、对外输出能力。2、**云服务需求扩张：**DeepSeek 成功证明海内外市场对高性能 AI 推理的强烈需求。已经有多家互联网大厂接入 DeepSeek，并提供价格等优惠。市场有望感知到并认可头部互联网公司云服务能力和未来增长空间，有望逐步提升云业务估值水平。3、**推理模型直接提升 AI Agent 能力：**OpenAI 上线 AI Agent，腾讯、阿里等互联网大厂提供 AI Agent 服务。AI Agent 仍处于早期，具备一定不可预测性，且依赖基座模型表现。4、**提升 AI 赋能广告能力：**广告作为互联网重要的高毛利变现板块，AI 赋能空间理想。有望逐步加深 AI 自动化广告的渗透，将大模型更深度地用于分析用户在大厂生态上的行为，精准总结画像进而优化广告定向投放，促进广告点击率等指标增长，进而提升广告业务收入和利润水平的增长率。

**投资建议：**复盘 DeepSeek-R1 在业界和大众间的“破圈”，关注投资主线：1) R1 等科技成果催化下中概资产价值重估。2) 技术进展等方面领先的中国大模型，推荐：阿里巴巴-W、腾讯控股、快手-W、百度集团-SW。3) 通过云提供 DeepSeek 服务实现收入增长、估值提振，推荐：阿里巴巴-W。4) 模型增强，提升 AI 广告渗透，推荐：腾讯控股、快手-W、哔哩哔哩、百度集团-SW。

**风险提示：**AI 技术研发和产品迭代不及预期；AI 行业竞争加剧风险；商业化进展不及预期风险；国内外政策风险。

## 互联网传媒

## 买入（维持）

## 作者

分析师：付天姿

执业证书编号：S0930517040002

021-52523692

futz@ebsecn.com

分析师：赵越

执业证书编号：S0930524020001

021-52523690

zhaoyue1@ebsecn.com

分析师：姜浩

执业证书编号：S0930522010001

021-52523680

jianghao@ebsecn.com

## 行业与纳斯达克指数对比图



资料来源：

## 相关研报

火山引擎冬季 FORCE 原动力大会召开，梳理字节 AI 全产业链——AIGC 行业跟踪报告（四十四）（2024-12-22）

掌趣 AI 游戏创作平台前瞻视频发布，AI 应用勤更新激发市场情绪势能——AIGC 行业跟踪报告（三十七）（2023-12-07）

关注 AIGC+游戏潜在催化：ChinaJoy2023 有何亮点？——AIGC 行业跟踪报告（十八）（2023-07-27）

游戏板块后续还有哪些潜在催化剂？——AIGC 行业跟踪报告（十六）（2023-7-20）

生成式 AI 管理办法落地，利于模型及应用加速发展——AIGC 行业跟踪报告（十五）（2023-7-13）

人工智能大会多场论坛聚焦游戏，AI 结合游戏进一步被市场认知——AIGC 行业跟踪报告（十四）（2023-07-11）

网易《逆水寒》手游初上线表现优异，AI 游戏迎关键玩法革新——AIGC 行业跟踪报告（十二）（2023-07-03）

# 并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

# 目 录

|                                                      |           |
|------------------------------------------------------|-----------|
| <b>1、DeepSeek-R1 “破圈” 拆解：突破性解决业界难题，全面开源利于传播.....</b> | <b>5</b>  |
| 1.1 原因 1：强大性能领跑全球，评测、实际体验赶超 OpenAI 推理模型 o1 .....     | 5         |
| 1.2 原因 2：多项工作实现算法和工程上的实质性突破，解决困扰行业的难题 .....          | 7         |
| 1.3 原因 3：全面开源并推出免费 C 端产品，使得优质 AI 推理体验快速扩散.....       | 9         |
| <b>2、互联网大厂大模型：持续迭代参与竞争，阿里通义性能比肩 DeepSeek.....</b>    | <b>10</b> |
| 2.1 阿里巴巴：基座模型、深度推理模型进展稳居第一梯队.....                    | 10        |
| 2.2 腾讯：基座模型采取跟随战略稳健追赶，组织架构调整聚焦应用结合 .....             | 13        |
| 2.3 百度：文心最早上线经多次迭代，期待 25 年下一代模型能力提升.....             | 16        |
| 2.4 快手：可灵模型专注文生视频领域居业界领先 .....                       | 17        |
| 2.5 美团：MobileVLM V2 模型在移动设备环境具备优势，或仍在推进自研模型 .....    | 19        |
| 2.6 哔哩哔哩：Index 自研模型在角色扮演、长文本等方面表现不俗.....             | 20        |
| <b>3、DeepSeek “破圈” 后，对互联网大厂 AI 应用有何价值？ .....</b>     | <b>21</b> |
| 3.1 大模型能力：R1 是起点不是终点，技术突破有望启发大厂改进模型.....             | 21        |
| 3.2 云服务：大厂已广泛支持 DeepSeek 模型，模型加速迭代有望提升云需求.....       | 22        |
| 3.3 AI Agent：优质推理模型带来能力提升，大厂 AI Agent&行业应用有望渗透 ..... | 23        |
| 3.4 AI+广告：AIGC 重塑营销链条，更强的模型效果提升广告自动化能力 .....         | 25        |
| <b>4、投资建议.....</b>                                   | <b>27</b> |
| <b>5、风险提示.....</b>                                   | <b>28</b> |

## 图目录

|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| 图 1: DeepSeek-R1 模型在多个基准测试中的表现超越 OpenAI-o1 模型的两个版本 mini 和 0912.....                    | 5  |
| 图 2: 海外 AI 社区对 DeepSeek-R1 的物理测试具有高评价 .....                                            | 7  |
| 图 3: DeepSeek 主力模型持续迭代性能逐步提升直至接近最优模型 .....                                             | 8  |
| 图 4: DeepSeek-R1 模型的开发过程.....                                                          | 9  |
| 图 5: AIGC APP 行业 TOP5 APP DAU 趋势 .....                                                 | 10 |
| 图 6: 超级产品增长 1 亿用户所用的时间（部分） .....                                                       | 10 |
| 图 7: Qwen2.5-Max 指令模型在基准测试中成绩优异.....                                                   | 11 |
| 图 8: Qwen2.5-Max 基座模型在基准测试中展现优势.....                                                   | 11 |
| 图 9: Qwen2.5-Max 一句话生成代码及可视化演示 .....                                                   | 11 |
| 图 10: Qwen2.5-Max 一句话生成扫雷小游戏演示.....                                                    | 11 |
| 图 11: QwQ-32B-Preview 在数学和编程等领域基准集中获得能力的提升 .....                                       | 12 |
| 图 12: Chatbot Arena 开源大模型榜单.....                                                       | 12 |
| 图 13: DeepSeek 采用 Qwen 开源模型蒸馏多个小模型 .....                                               | 12 |
| 图 14: 24M12 MAU TOP10 综合类 AI 原生 App.....                                               | 14 |
| 图 15: 腾讯会议 AI 小助手 Pro 功能升级.....                                                        | 14 |
| 图 16: 混元 Hunyuan-Large 模型在 MATH、HumanEval 等测评集效果好于 Llama3.1-405B 及 DeepSeek-V2.5 ..... | 14 |
| 图 17: 混元 3D 支持自主设计细致的 3D 生成工作流 .....                                                   | 15 |
| 图 18: 混元 3D 几何模型生成可视化比较.....                                                           | 15 |
| 图 19: 百舸产品架构.....                                                                      | 16 |
| 图 20: 千帆三层架构.....                                                                      | 16 |
| 图 21: 可灵 1.6 图生视频稳定性提升.....                                                            | 18 |
| 图 22: 可灵 1.6 图生视频人物运动表演加强.....                                                         | 18 |
| 图 23: MobileVLM V2 在速度和准确性上均有提升 .....                                                  | 19 |
| 图 24: Index-1.9B-32K 与 GPT-4、Qwen2 等模型长文本能力对比 .....                                    | 20 |
| 图 25: Index-1.9B-Character 在角色扮演林黛玉中的表现.....                                           | 20 |
| 图 26: 三种 Scaling Law（预训练、后训练和在线推理）示意.....                                              | 21 |
| 图 27: DeepSeek 研究过程中对 Qwen、Llama 蒸馏模型与非蒸馏模型的基准比较 .....                                 | 22 |
| 图 28: DeepSeek-R1 推理成本和主流模型对比.....                                                     | 24 |
| 图 29: 腾讯元器创建 AI Agent 页面.....                                                          | 25 |
| 图 30: 腾讯元器特色功能和分发渠道 .....                                                              | 25 |
| 图 31: 提出自动竞价策略示例 .....                                                                 | 26 |
| 图 32: 谷歌 AI 概要广告.....                                                                  | 26 |
| 图 33: 百度城市名片智能体示例 .....                                                                | 27 |
| 图 34: 腾讯广告妙思具备文生视频等全面广告创意素材生成能力 .....                                                  | 27 |

表目录

表 1: Chatbot Arena 榜单中 DeepSeek-R1 居前列 (2025-02-09) ..... 6

表 2: 科技媒体 arstechnica 对 DeepSeek-R1 与 OpenAI-o1&o1-Pro 分别提问并对答案进行评测 ..... 6

表 3: DeepSeek 通过模型迭代逐步引入多项创新性技术突破 ..... 8

表 4: 腾讯 AI 大模型主要布局团队 ..... 13

表 5: 23M5-24M10 百度文心大模型迭代&相关产品发布时间线梳理 ..... 17

表 6: AGI-Eval 评测社区 24 年 12 月多模态模型 (文生视频模型) 榜单排行 ..... 19

表 7: 主要互联网大厂提供 DeepSeek-R1 调用服务价格 ..... 23

表 8: AgentOps.ai 25 年 AI Agent 图谱 (25 年 1 月) ..... 25

# 1、DeepSeek-R1 “破圈” 拆解：突破性解决业界难题，全面开源利于传播

## 1.1 原因 1：强大性能领跑全球，评测、实际体验赶超 OpenAI 推理模型 o1

DeepSeek 的旗舰推理模型 R1 在多个基准测试中的表现超越当下全球 AI 行业领先的推理模型 OpenAI-o1。根据 DeepSeek-R1 公开的技术报告，经过额外的 SFT 阶段和进一步的 RL 训练完善后的 R1，在 AIME 2024、MATH-500、LiveCode Bench、CodeForces 等多个数学、编程测试集中获得超越 OpenAI 的 o1 系列的分数，仅在考察物理化学生物的 GPQA Diamond 数据集上逊色于 OpenAI-o1-0912。

图 1：DeepSeek-R1 模型在多个基准测试中的表现超越 OpenAI-o1 模型的两个版本 mini 和 0912

| Model            | Math benchmarks |         |          | Bio, physics & chemistry | Code benchmarks |            |
|------------------|-----------------|---------|----------|--------------------------|-----------------|------------|
|                  | AIME 2024       |         | MATH-500 | GPQA Diamond             | LiveCode Bench  | CodeForces |
|                  | pass@1          | cons@64 | pass@1   | pass@1                   | pass@1          | rating     |
| OpenAI-o1-mini   | 63.6            | 80.0    | 90.0     | 60.0                     | 53.8            | 1820       |
| OpenAI-o1-0912   | 74.4            | 83.3    | 94.8     | 77.3                     | 63.4            | 1843       |
| DeepSeek-R1-Zero | 71.0            | 86.7    | 95.9     | 73.3                     | 50.0            | 1444       |
| DeepSeek-R1      | 79.8            |         | 97.3     | 71.5                     | 65.9            | 2029       |

Higher is better

RL only →

SFT + RL →

资料来源：Ahead of AI 博客，DeepSeek-R1 技术报告

DeepSeek-R1 在开发人员和使用者中收获高评价，其在 Chatbot Arena 榜单中位居前列，超过 OpenAI-o1。Chatbot Arena 是一个基于人类偏好评估 LLM 的开放平台，其方法采用成对比较方法，用户只需投票比较两个模型响应并投票选出更好的一个，平台通过众包利用来自不同用户群的输入，截至 2025 年 2 月 9 日，平台共收集到超过 260 万次用户的投票。尽管 DeepSeek-R1 上线时间较晚，尚未收集到足够多的投票次数（共 4193 次，前十名的模型中最少），但仍获得 1361 分的 Arena Elo 分数，超过 OpenAI-o1，仅次于 Gemini 的两款模型和最新版的 ChatGPT-4o。

表 1：Chatbot Arena 榜单中 DeepSeek-R1 居前列（2025-02-09）

| 模型                                  | Arena Elo 分数 | 编程 Elo 分数 | 投票次数  | 机构       |
|-------------------------------------|--------------|-----------|-------|----------|
| Gemini-2.0-Flash-Thinking-Exp-01-21 | 1384         | 1364      | 11462 | Google   |
| Gemini-2.0-Pro-Exp-02-05            | 1379         | 1373      | 9385  | Google   |
| ChatGPT-4o-latest (2024-11-20)      | 1365         | 1351      | 39649 | OpenAI   |
| DeepSeek-R1                         | 1361         | 1363      | 4193  | DeepSeek |
| Gemini-2.0-Flash-001                | 1355         | 1349      | 7264  | Google   |
| o1-2024-12-17                       | 1351         | 1363      | 13416 | OpenAI   |
| Qwen2.5-Max                         | 1332         | 1337      | 5459  | Alibaba  |
| DeepSeek-V3                         | 1317         | 1317      | 16463 | DeepSeek |
| Gemini-2.0-Flash-Lite-Preview-02-05 | 1309         | 1317      | 6638  | Google   |
| o3-mini                             | 1309         | 1362      | 6284  | OpenAI   |

资料来源：Chatbot Arena，光大证券研究所

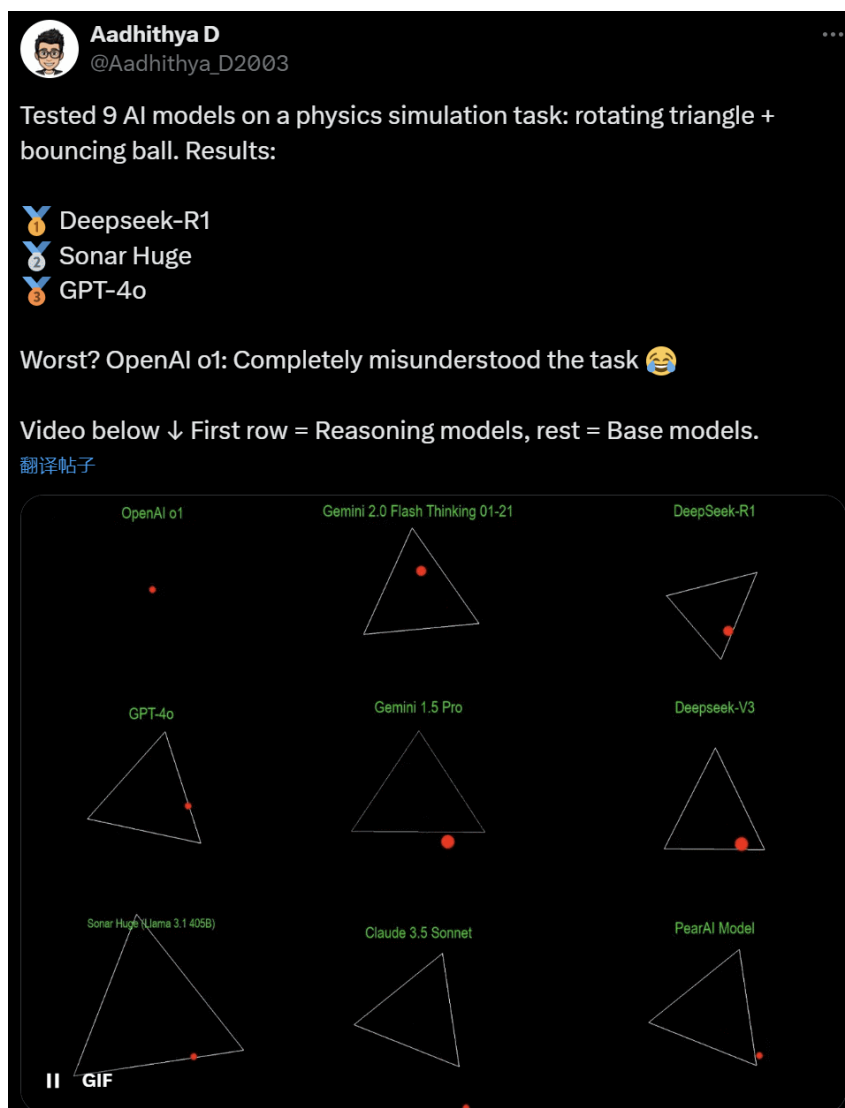
DeepSeek-R1 在实际用户体验中收获不亚于 OpenAI-o1 系列的广泛好评，包括英文日常问答、物理测试等。1) 海外科技媒体 arstechnica 资深编辑对 DeepSeek-R1 与 OpenAI-o1 和 OpenAI-o1-Pro 进行评测，该评测更侧重于模拟英文用户可能提出的日常问题，测试中所用的提问涵盖创意写作、数学、指令遵循等领域以及部分设计得更加复杂、要求更高且更严谨的提问。该团队考虑模型回答的正确性及一些主观质量因素，最终评测结果 DeepSeek-R1 在多个提问中返回优于 OpenAI-o1 系列的回复。2) 根据机器之心报道，CoreView CTO Ivan Fioravanti 等海外 AI 社区人士评测称，DeepSeek-R1 在编写 Python 脚本，模拟小球在一个旋转的形状中弹跳的物理测试中，收获优于 OpenAI GPT-4o 等其他大模型的表现。

表 2：科技媒体 arstechnica 对 DeepSeek-R1 与 OpenAI-o1&o1-Pro 分别提问并对答案进行评测

| 序号 | 提问                                                                                                                              | 媒体评测获胜者                     |
|----|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------|
| 1  | 写五个原创的老爸笑话（轻松、无害，还有点「冷」的笑话）。                                                                                                    | DeepSeek-R1                 |
| 2  | 写一篇关于亚伯拉罕·林肯发明篮球的两段创意故事。                                                                                                        | DeepSeek-R1                 |
| 3  | 写一段短文，其中每句话的第二个字母拼出单词「CODE」。这段文字应显得自然，不要明显暴露这一模式。                                                                               | ChatGPT-o1-Pro              |
| 4  | 如果 Magenta 这个城镇不存在，这种颜色还会被称为「品红」（magenta）吗？                                                                                     | ChatGPT-o1-Pro              |
| 5  | 第 10 亿个质数是多少？                                                                                                                   | DeepSeek-R1                 |
| 6  | 我需要你帮我制定一个时间表，基于以下几点：我的飞机早上 6:30 起飞、需要在起飞前 1 小时到达机场、去机场需要 45 分钟、我需要 1 小时来穿衣和吃早餐。<br>请一步一步考虑，告诉我应该几点起床，什么时候出发，这样才能准时赶上 6:30 的航班。 | DeepSeek-R1                 |
| 7  | 在我的厨房里，有一张桌子，上面放着一个杯子，杯子里有一个球。我把杯子移到了卧室的床上，并将杯子倒过来。然后，我再次拿起杯子，移到了主房间。现在，球在哪里？                                                   | 并列                          |
| 8  | 请提供一个包含 10 个自然数的列表，要求满足：至少有一个是质数，至少 6 个是奇数，至少 2 个是 2 的幂次方，并且这 10 个数的总位数不少于 25 位。                                                | ChatGPT-o1 和 ChatGPT-o1-Pro |

资料来源：机器之心，arstechnica，光大证券研究所

图 2：海外 AI 社区对 DeepSeek-R1 的物理测试具有高评价



资料来源：机器之心，X

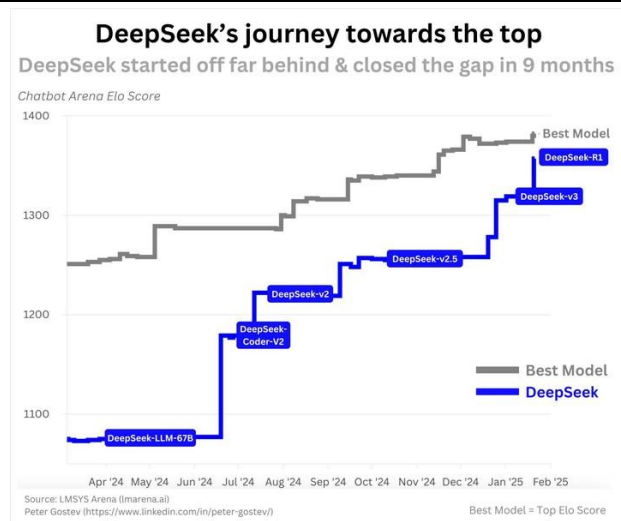
## 1.2 原因 2：多项工作实现算法和工程上的实质性突破，解决困扰行业的难题

DeepSeek-R1 经历多代模型的迭代，逐步实现在 GRPO 算法、DeepSeekMoE 架构、MLA 机制、FP8 精度、MTP 方法等多方位的突破。DeepSeek 的首个开源模型 DeepSeek-Coder，于 2023 年 11 月发布，当时是业界领先的代码大模型。随后一年多时间，DeepSeek 逐步推出 DeepSeekMath、DeepSeek-V2、DeepSeek-V3 等多款大模型，不仅提升模型性能，更引入 GRPO 算法、DeepSeekMoE 架构、MLA 机制、FP8 精度、MTP 方法等多项切实提升模型推理能力、训练效率的创新点，并在技术报告中详细阐述，得到 AI 业界的广泛讨论和认可。

**表 3：DeepSeek 通过模型迭代逐步引入多项创新性技术突破**

| 创新性技术突破        | 引入模型         | 发布时间        | 简介                                                           |
|----------------|--------------|-------------|--------------------------------------------------------------|
| GRPO 算法        | DeepSeekMath | 2024 年 2 月  | 群组相对策略优化（GRPO）算法，这是对经典 PPO 算法的创新改进，不仅增强了模型的数学推理能力，还优化了内存使用效率 |
| DeepSeekMoE 架构 | DeepSeek-V2  | 2024 年 5 月  | 通过细粒度的专家分割和共享专家隔离，DeepSeekMoE 与主流的 MoE 架构相比，实现了更高的专家专业化和性能。  |
| MLA 机制         | DeepSeek-V2  | 2024 年 5 月  | 创新多头潜在注意力（MLA）机制，性能优于传统的 MHA，但需要的 KV 缓存量要少得多                 |
| FP8 精度         | DeepSeek-V3  | 2024 年 12 月 | 在极大规模模型上验证了 FP8 训练的有效性，通过支持 FP8 计算和存储，实现加速训练和减少 GPU 内存使用。    |
| MTP 方法         | DeepSeek-V3  | 2024 年 12 月 | MTP 目标使训练信号更加密集，并可能提高数据效率。MTP 可以使模型预先规划其表示，以便更好地预测后续的 token。 |

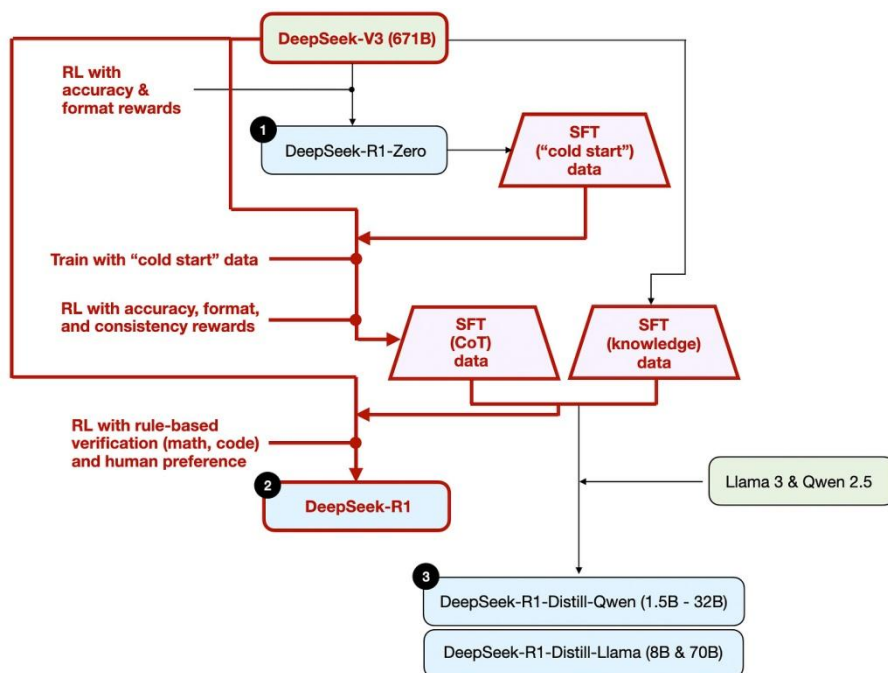
资料来源：csdn 博客，DeepSeek 各模型技术报告，光大证券研究所

**图 3：DeepSeek 主力模型持续迭代性能逐步提升直至接近最优模型**

资料来源：爱范儿，LMSYS Arena

DeepSeek-R1 是行业首个验证在后训练 (Post-Training) 时使用强化学习 (RL) 让千亿参数的模型获得推理能力的研究，切实解决行业难题。区别于之前业界主流的探索方向是过程奖励模型 (PRM) 和蒙特卡洛树搜索 (MCTS)，DeepSeek 跳过指令优化的监督微调 (SFT) 阶段直接用强化学习得到 DeepSeek-R1-Zero 模型，且只采用准确性和格式两种类型的奖励，使得模型生成推理轨迹并产生推理能力。DeepSeek-R1 指出一条除了大幅提升训练数据量之外获得效果更好模型的技术路线。

图 4: DeepSeek-R1 模型的开发过程



资料来源: Ahead of AI 博客

### 1.3 原因 3: 全面开源并推出免费 C 端产品, 使得优质 AI 推理体验快速扩散

DeepSeek 采用完全开源策略, 降低 C 端用户使用门槛, 促进 AI 开发者社区的协作生态。

相较于闭源且收费较高的 OpenAI-o1, 1) 通过开源并在技术报告中详细公布技术进展和模型训练思路, DeepSeek 吸引大量海内外开发者和研究人员的关注, 使得其作为中国模型首次受到海外 AI 科技界全面推崇认可。2) 免费使用的 DeepSeek App 成为大部分中国乃至全球用户首次体验的优质 AI 推理模型, 用户量实现快速增长。根据 Questmobile 数据, DeepSeek 在 25 年 1 月 28 日的日活跃用户数首次超越豆包, 随后在 2 月 1 日突破 3000 万大关, 成为史上最快达成这一里程碑的应用。根据 AI 产品榜数据, 25 年 1 月 DeepSeek 用户增长达 1.25 亿 (含网站 (Web)、应用 (App) 累加不去重)。其中, 80% 以上用户来自 1 月最后一周, 即 DeepSeek 在几乎没有任何广告投放情况下实现 7 天完成 1 亿用户增长。

图 5: AIGC APP 行业 TOP5 APP DAU 趋势



资料来源: Questmobile

图 6: 超级产品增长 1 亿用户所用的时间 (部分)



## 2、互联网大厂大模型：持续迭代参与竞争，阿里通义性能比肩 DeepSeek

### 2.1 阿里巴巴：基座模型、深度推理模型进展稳居第一梯队

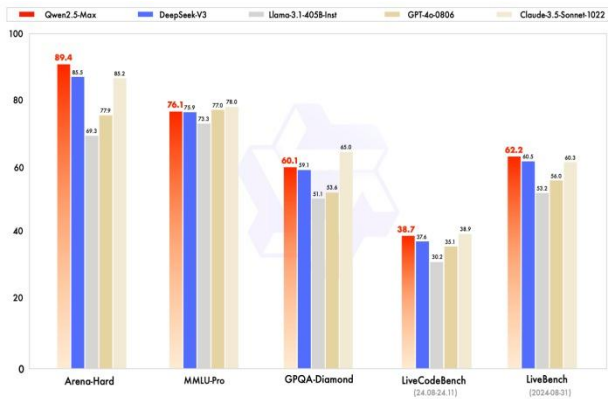
阿里旗下最新旗舰模型 Qwen2.5-Max 在指令模型、基座模型的指标对比中，均已赶超业界领先的模型。阿里通义于 25 年 1 月发布最新 Qwen2.5-Max 模型，其为通义千问系列效果最好的模型。

根据通义千问披露：1) 指令模型（即我们平常使用的可以直接对话的模型）对比，在 Arena-Hard、LiveBench、LiveCodeBench 和 GPQA-Diamond 等基准测试中，Qwen2.5-Max 的表现超越 DeepSeek-V3。同时在 MMLU-Pro 等其他评估中也展现出具备竞争力的成绩。

2) 基座模型对比中，Qwen2.5-Max 在 MMLU (大规模多任务语言理解)、MATH、BBH 等多项测试中均展现出相对上一代 Qwen2.5-72B 的大幅提升，以及相对 DeepSeek-V3 的超越。

3) 尽管并未进一步披露在算法技术、工程上的具体细节，但 Qwen2.5-Max 同样为超大规模的 MoE 模型，使用超过 20 万亿 token 的预训练数据及精心设计的后训练方案进行训练。Qwen2.5-Max 和 DeepSeek-V3 同样实现 AI 业界对训练超大规模 MoE 模型的突破。

图 7: Qwen2.5-Max 指令模型在基准测试中成绩优异



资料来源: 通义千问 github

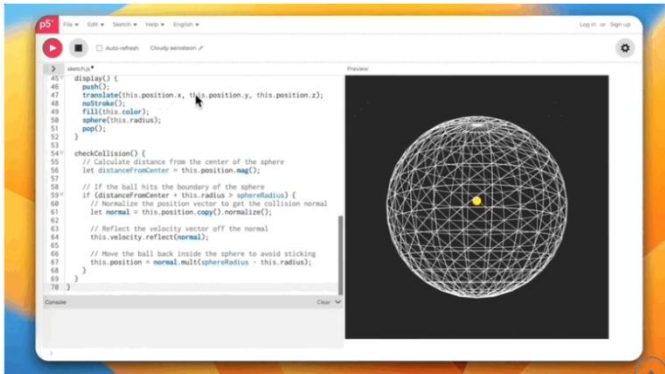
图 8: Qwen2.5-Max 基座模型在基准测试中展现优势

|           | Qwen2.5-Max | Qwen2.5-72B | DeepSeek-V3 | Llama3.1-405B |
|-----------|-------------|-------------|-------------|---------------|
| MMLU      | 87.9        | 86.1        | 87.1        | 85.2          |
| MMLU-Pro  | 69.0        | 58.1        | 64.4        | 61.6          |
| BBH       | 89.3        | 86.3        | 87.5        | 85.9          |
| C-Eval    | 92.2        | 90.7        | 90.1        | 72.5          |
| CMMLU     | 91.9        | 89.9        | 88.8        | 73.7          |
| HumanEval | 73.2        | 64.6        | 65.2        | 61.0          |
| MBPP      | 80.6        | 72.6        | 75.4        | 73.0          |
| CRUX-I    | 70.1        | 60.9        | 67.3        | 58.5          |
| CRUX-O    | 79.1        | 66.6        | 69.8        | 59.9          |
| GSMBK     | 94.5        | 91.5        | 89.3        | 89.0          |
| MATH      | 68.5        | 62.1        | 61.6        | 53.8          |

资料来源: 通义千问 github

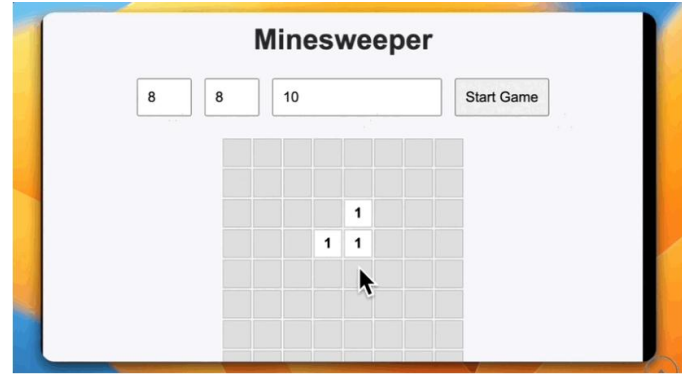
Qwen2.5-Max 代码编写等各项能力、实际应用体验均得到提升, 已在 Qwen Chat 中上线, 整体接入阿里云服务 API。1) Qwen2.5-Max 的代码编写与理解能力、逻辑能力、多语言能力显著提升, 回复风格面向人类偏好进行大幅调整, 模型回复详实程度和格式清晰度明显改善, 内容创作、JSON 格式遵循、角色扮演能力定向提升。2) Qwen2.5-Max 具备联网搜索功能, 输出的每句话来源出处都有标注, 整体运行也很丝滑。代码能力上, Qwen2.5-Max 能够帮助用户完成各种可视化创作, 一句话生成代码及建模; 也有 Artifacts 功能, 一句话能开发各种小应用、小游戏。

图 9: Qwen2.5-Max 一句话生成代码及可视化演示



资料来源: 量子位

图 10: Qwen2.5-Max 一句话生成扫雷小游戏演示



资料来源: 量子位

阿里旗下最新实验性研究推理模型在数学和编程等领域已取得显著进步, 期待 Qwen2.5-Max 新模型赋能、DeepSeek-R1 开源后的技术启示, QwQ-32B 正式版带来突破。阿里通义于 24 年 11 月发布 QwQ-32B-Preview 实验性研究模型, 专注于增强 AI 推理能力。

作为预览版本, 根据通义千问披露: 1) QwQ-32B-Preview 和 OpenAI-o1-preview 和 OpenAI-o1-mini 在 GPQA (科学推理)、AIME、MATH-500 (数学) 以及 LiveCodeBench (代码) 四个数据集中各有胜负, 但整体水平比较接近。而相比 GPT-4o、Claude 3.5 Sonnet 和 Qwen2.5, 具备比较明显的领先优势。

2) 作为预览版本, QwQ-32B-Preview 仍存在语言切换问题、推理循环、安全性考虑和能力差异等问题。而 DeepSeek-R1 已解决部分此局限性, 其多方面能力超过 OpenAI-o1-mini, 在奖励函数设计中, 重点对语言一致性进行的要求。DeepSeek-R1 的模型训练思路有望成为 QwQ-32B 的借鉴, 同时在 25 年 2 月发布的更强大的 Qwen2.5-Max 则有望成为 QwQ-32B 训练的基石。

图 11: QwQ-32B-Preview 在数学和编程等领域基准集中获得能力的提升

|                                  | QwQ<br>32B-preview | OpenAI<br>o1-preview | OpenAI<br>o1-mini | GPT-4o | Claude3.5<br>Sonnet | Qwen2.5-72B<br>Instruct |
|----------------------------------|--------------------|----------------------|-------------------|--------|---------------------|-------------------------|
| GPQA<br>Pass@1                   | 65.2               | 72.3                 | 60.0              | 53.6   | 65.0                | 49.0                    |
| AIME<br>Pass@1                   | 50.0               | 44.6                 | 56.7              | 9.3    | 16.0                | 23.3                    |
| MATH-500<br>Pass@1               | 90.6               | 85.5                 | 90.0              | 76.6   | 78.3                | 82.6                    |
| LiveCodeBench<br>2024-08-2024.11 | 50.0               | 53.6                 | 58.0              | 33.4   | 36.3                | 30.4                    |

资料来源: 通义千问 github

阿里通义同样是模型开源的支持和践行者, 其在开源大模型中性能和开发者参与度均居领先地位。1) Qwen 系列模型中, 除了旗舰模型闭源商用外, 其余所有模型都在走开源路线。截至 2025 年 2 月 9 日, Chatbot Arena 排名中, Qwen2.5-72B-Instruct 居第三位, 优于 Llama-3.3-70B-Instruct。

2) Qwen 系列的特点是开源模型提供全尺寸开源模型。和 DeepSeek-V3/R1 开源的 671B 超大模型不同, Qwen 开源模型参数量覆盖小到手机也能运行的 1.5B, 大到 110B, 基本上能覆盖开源社区的绝大多数需求, 因而在全球开源社区中影响力很大, AI 业界非常多的研究工作都是以 Qwen 为基础模型开展的。25 年 2 月 10 日, 全球最大 AI 开源社区 Huggingface 发布了最新的开源大模型榜单, 其中排名前十的开源大模型, 都为基于阿里通义千问开源模型二次训练的衍生模型。DeepSeek-R1 基于 Qwen 2.5 模型 (参数个数 1.5B 到 32B) 蒸馏多个小模型, 提供更具效率的版本, 并做案例探索研究。

图 12: Chatbot Arena 开源大模型榜单

| Best Open LM           |             |        |           |
|------------------------|-------------|--------|-----------|
| Model ‡                | Arena Elo ‡ | MMLU ‡ | License ‡ |
| DeepSeek-R1            | 1361        | 90.8   | MIT       |
| DeepSeek-V3            | 1317        | 88.5   | DeepSeek  |
| Qwen2.5-72B-Instruct   | 1257        | 86.8   | Qwen      |
| Llama-3.3-70B-Instruct | 1255        | 86     | Llama 3.3 |

资料来源: Chatbot Arena

图 13: DeepSeek 采用 Qwen 开源模型蒸馏多个小模型

| Model                         | AIME 2024 |         | MATH-500 | GPQA<br>Diamond | LiveCode<br>Bench | CodeForces |
|-------------------------------|-----------|---------|----------|-----------------|-------------------|------------|
|                               | pass@1    | cons@64 | pass@1   | pass@1          | pass@1            | rating     |
| GPT-4o-0513                   | 9.3       | 13.4    | 74.6     | 49.9            | 32.9              | 759        |
| Claude-3.5-Sonnet-1022        | 16.0      | 26.7    | 78.3     | 65.0            | 38.9              | 717        |
| OpenAI-o1-mini                | 63.6      | 80.0    | 90.0     | 60.0            | 53.8              | 1820       |
| QwQ-32B-Preview               | 50.0      | 60.0    | 90.6     | 54.5            | 41.9              | 1316       |
| DeepSeek-R1-Distill-Qwen-1.5B | 28.9      | 52.7    | 83.9     | 33.8            | 16.9              | 954        |
| DeepSeek-R1-Distill-Qwen-7B   | 55.5      | 83.3    | 92.8     | 49.1            | 37.6              | 1189       |
| DeepSeek-R1-Distill-Qwen-14B  | 69.7      | 80.0    | 93.9     | 59.1            | 53.1              | 1481       |
| DeepSeek-R1-Distill-Qwen-32B  | 72.6      | 83.3    | 94.3     | 62.1            | 57.2              | 1691       |
| DeepSeek-R1-Distill-Llama-8B  | 50.4      | 80.0    | 89.1     | 49.0            | 39.6              | 1205       |
| DeepSeek-R1-Distill-Llama-70B | 70.0      | 86.7    | 94.5     | 65.2            | 57.5              | 1633       |
| DeepSeek-R1-Zero              | 71.0      |         | 95.9     | 73.3            | 50.0              | 1444       |
| DeepSeek-R1                   | 79.8      |         | 97.3     | 71.5            | 65.9              | 2029       |

资料来源: Ahead of AI 博客, DeepSeek-R1 技术报告

## 2.2 腾讯：基座模型采取跟随战略稳健追赶，组织架构调整聚焦应用结合

腾讯在 AI 大模型中的团队布局较为分散，持续关注人员、资源配置重点。包括：

- 1) **混元大模型团队**，旗下产品包括 23 年 9 月正式上线的混元系列大模型，及基于混元大模型及搜索引擎驱动的 AI 智能助手“元宝”和 AI 智能体开放平台“元器”。
- 2) **腾讯 AI Lab 团队**，早在 16 年 4 月成立，其基础研究方向包括计算机视觉、语音技术、自然语言处理和机器学习，应用探索结合了腾讯场景与业务优势，聚焦于游戏、数字人、内容和社交 AI 四类。
- 3) **腾讯云 AI 团队**，23 年 6 月早于混元发布行业大模型，并发布面向 B 端客户的腾讯云 MaaS 服务解决方案，腾讯云板块基础设施持续支持腾讯内部模型研发和应用探索。
- 4) **其他应用团队**，腾讯内每个事业群内均在探索大模型的产品化落地场景，包括微信、QQ、输入法、浏览器等产品都将推出 AI 智能体，游戏、微信读书、腾讯视频等产品也将基于混元做更多 AI 探索。**以微信 AI 团队为例**，24M1 微信公开课 PRO 分享微信对话开放平台的更新，其能够帮助开发者和商家快速搭建一个零成本、低门槛，满足自身业务需要的对话机器人。

表 4：腾讯 AI 大模型主要布局团队

| 腾讯大模型布局团队 | 归属事业群 | 简介                                                           |
|-----------|-------|--------------------------------------------------------------|
| 混元大模型     | TEG   | 23 年 9 月上线；截至 24 年底，相继开源旗下文生文、文生图、3D 生成大模型和视频生成大模型。          |
| 腾讯元宝      | CSIG  | 24 年 5 月上线；基于混元大模型及搜索引擎驱动的 AI 智能助手。                          |
| 腾讯元器      | TEG   | 24 年 5 月上线；混元大模型团队推出的 AI 智能体开放平台。                            |
| 腾讯 AI Lab | TEG   | 16 年 4 月成立；基础研究方向计算机视觉、语音技术、自然语言处理和机器学习，应用探索游戏、数字人、内容和社交 AI。 |
| 腾讯云       | CSIG  | 23 年 6 月发布行业大模型，并发布面向 B 端客户的腾讯云 MaaS 服务解决方案。                 |
| 微信 AI     | WXG   | 24M1 微信公开课 PRO 分享微信对话开放平台，帮助开发者和商家快速搭建 AI 对话机器人。             |

资料来源：腾讯混元官网，腾讯 AI Lab，光子星球，极客公园，光大证券研究所

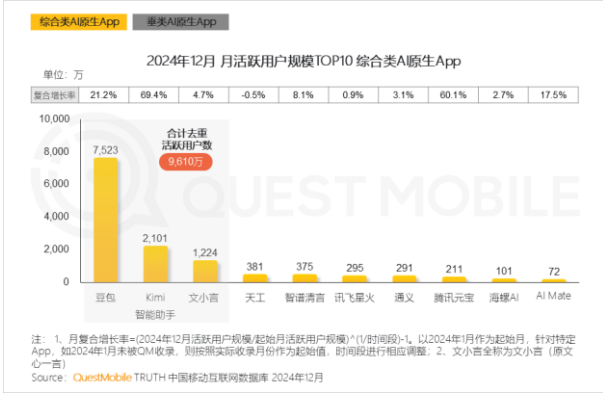
**组织架构调整，元宝并入 CSIG，或彰显腾讯 AI 战略更加清晰，从技术到重视产品体验，从“后手入场”到全面探索 AI 应用的转变。**25 年初，腾讯 AI 助手应用“元宝”完成组织调整，产品团队从 TEG 事业群（技术工程事业群）调整至 CSIG（云与智慧产业事业群）。调整后，“元宝”应用将交由腾讯会议负责人吴祖榕负责，主要负责元宝的产品能力建设和体验优化。

我们认为：1) **元宝从技术部门的剥离或显示出腾讯 AI 战略从技术到重视产品体验的转变。**吴祖榕是一个具有丰富 ToB 经验且能兼顾 C 端产品体验的负责人，元宝有望复制腾讯会议经验，以 C 端体验为入口，连接和服务企业客户，进而探索 AI 商业化。腾讯会议基于混元大模型已经推出智能录制、智能生成摘要总结、腾讯会议 AI 助手等功能。

2) **元宝 App 等腾讯 AI 产品 24 年整体进展相对保守，后续或有望在 AI 产品化方面转向积极。**腾讯公司一贯更注重长期维持优质的产品设计和用户体验，通常谨慎对待新技术新概念融入自身产品矩阵和社交体系，在元宝 App 的推广方面较为保守。根据 Questmobile 数据，24 年 12 月腾讯元宝 App MAU 211 万，明显低于字节豆包 App、百度文小言 App 等。根据第一财经等媒体报道，25 年 1 月腾讯集团年会中，董事会主席兼 CEO 马化腾表示，期望做腾讯混元的端到端语音交互落地。TEG 进行架构调整，将更聚焦做技术底座，产品化则希望其他

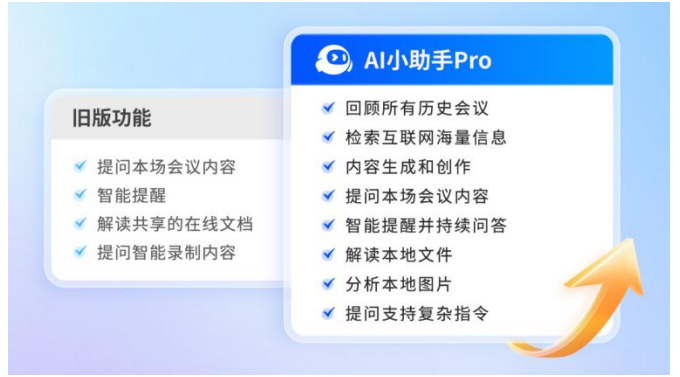
事业群一起推进。腾讯混元已经在跟腾讯会议、输入法、浏览器等结合，微信、QQ 都在推进智能体落地了，游戏也要全方位拥抱 AI。

图 14：24M12 MAU TOP10 综合类 AI 原生 App



资料来源：Questmobile

图 15：腾讯会议 AI 小助手 Pro 功能升级



资料来源：腾讯会议

基础大模型方面，腾讯混元最新开源 Hunyuan-Large 模型，模型效果整体赶超 Llama3.1-405B 及 DeepSeek-V2.5。根据混元披露：1) 腾讯混元 24 年 11 月发布的 Hunyuan-Large (Hunyuan-MoE-A52B) 模型，是当时业界已经开源的基于 Transformer 的最大 MoE 模型，拥有 389B 总参数和 52B 激活参数（对比 DeepSeek-V3 总参数量 671B，每个 Token 激活的参数量为 37B）。2) Hunyuan-Large 在 CMMLU、MMLU、CEval、MATH 等多学科综合评测集以及中英文 NLP 任务、代码和数学等维度取得理想成绩，在 MMLU、MATH、HumanEval 超越 Llama3.1-405B 及 DeepSeek-V2.5，在 ARC-C、GPQA\_diamond 不如 Llama3.1-405B，BBH、HellaSwag 不如 DeepSeek-V2.5，整体成绩略好于这两个当时领先的开源模型。

图 16：混元 Hunyuan-Large 模型在 MATH、HumanEval 等测评集效果好于 Llama3.1-405B 及 DeepSeek-V2.5

| Model                | LLama3.1<br>405B Inst. | LLama3.1<br>70B Inst. | Mixtral<br>8x22B Inst. | DeepSeek<br>V2.5 Chat | Hunyuan-Large<br>Inst. |
|----------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|
| MMLU                 | 87.3                   | 83.6                  | 77.8                   | 80.4                  | <b>89.9</b>            |
| BBH                  | -                      | -                     | 82.0                   | <b>87.1</b>           | 81.2                   |
| HellaSwag            | -                      | -                     | 86.0                   | <b>90.3</b>           | 88.5                   |
| ARC-C                | <b>96.9</b>            | 94.8                  | 91.5                   | 92.9                  | 94.6                   |
| DROP                 | -                      | -                     | 67.5                   | 79.5                  | <b>88.3</b>            |
| GPQA_diamond         | <b>50.7</b>            | 46.7                  | 38.4                   | 42.4                  | 42.4                   |
| MATH                 | 73.8                   | 68.0                  | 51.0                   | 74.7                  | <b>77.4</b>            |
| HumanEval            | 89.0                   | 80.5                  | 75.6                   | 89.0                  | <b>90.0</b>            |
| C-Eval               | -                      | -                     | 60.0                   | 79.9                  | <b>88.6</b>            |
| CMMLU                | -                      | -                     | 61.0                   | 79.5                  | <b>90.4</b>            |
| AlignBench           | -                      | -                     | 6.2                    | 8.0                   | <b>8.3</b>             |
| MT-Bench             | -                      | -                     | 8.1                    | 9.0                   | <b>9.4</b>             |
| IFEval strict-prompt | -                      | -                     | 71.2                   | -                     | <b>85.0</b>            |

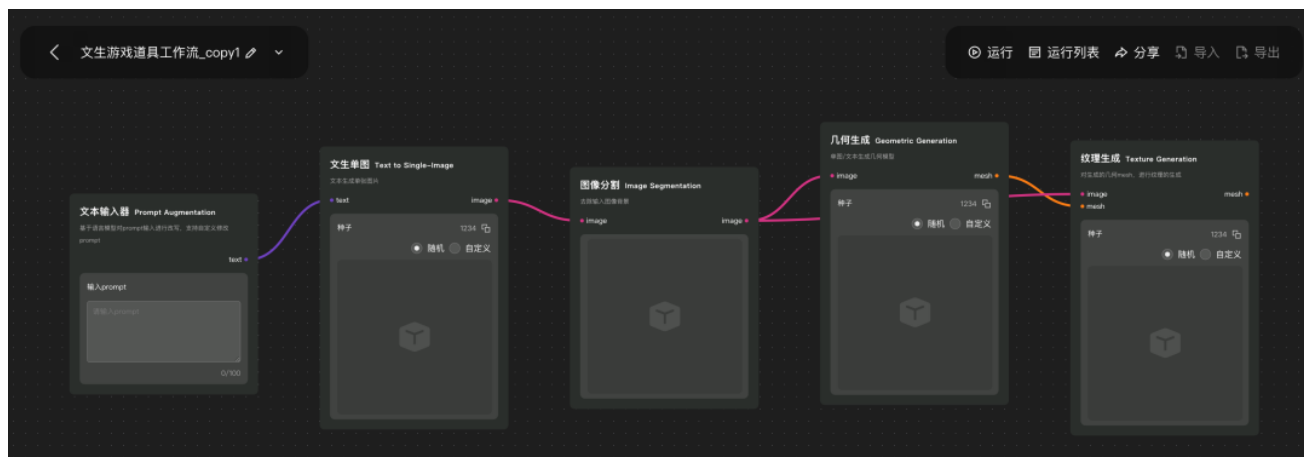
资料来源：量子位

腾讯混元在多模态方面具有较广布局和较多进展，探索 3D 生成、文生视频等领域，为内部赋能和行业进步打下基础。

1) 25 年 1 月，腾讯开源 3D 生成大模型 2.0 升级版本，上线业界首个一站式 3D 内容 AI 创作平台——混元 3D AI 创作引擎。作为创作者，可以用它输入文字、图

片一键生成高质量 3D 模型，并包含 3D 功能矩阵、3D 编辑、3D 生成工作流、创作素材库等多种功能。作为游戏开发、动画制作等领域专业创作者，还支持快速搭建 3D 生成工作流。

图 17：混元 3D 支持自主设计细致的 3D 生成工作流



资料来源：游戏葡萄

混元 3D AI 模型 2.0 版本再升级，通过几何、纹理解耦生成，几何结构更精细，纹理色彩更丰富。几何模型实现超高精度白模生成，媲美设计师手工建模水平。纹理模型则能对任意几何模型生成逼真纹理，支持文本/图像引导。

图 18：混元 3D 几何模型生成可视化比较



资料来源：量子位

2) 24 年 12 月，腾讯宣布旗下混元视频生成大模型（HunYuan-Video）开源，模型参数量 130 亿。该模型可供企业与个人开发者免费使用，目前已上线腾讯元宝 APP。HunYuan-Video 在文生视频多个方面都具有较高的质量，拥有包括超写实画质、原生镜头切换、高语义一致等特点。

## 2.3 百度：文心最早上线经多次迭代，期待 25 年下一代模型能力提升

百度自率先发布国产大模型文心以来，推出多次大模型迭代及相关产品发布，期待 25 年文心新版本面世。24 年 6 月，百度发布文心大模型 4.0 Turbo，大幅强化检索能力以改善幻觉问题，全网搜索、分析资料、等待大模型回复的速度得到明显提升。24 年 11 月，百度集团 CEO 李彦宏称文心的新版本面世，或在 25 年年初。

百度通过两大关键产品：大模型精调和应用开发平台的千帆，以及提供稳定高效算力服务的百舸平台，为企业提供全栈服务解决方案。1) 百舸：以 GPU 为核心搭建的异构计算平台，适合多模态大模型训练。百舸平台支持同一智算集群中混合使用不同厂商芯片，降低算力成本的同时，多芯混合训练任务的性能损失，控制在万卡性能损失 5%，已经是业界最高水平。2) 千帆：主打低门槛的模型平台，凭借模型开发层、模型服务层和应用开发层三层架构，满足多样化的现实需求。在模型开发层，千帆提供全流程工具；在模型服务层，可直接调用多模态能力；在应用开发层，千帆帮助企业用多模态能力改造业务。

图 19：百舸产品架构



资料来源：百度百舸

图 20：千帆三层架构



资料来源：机器之心

表 5：23M5-24M10 百度文心大模型迭代&amp;相关产品发布时间线梳理

| 时间    | 迭代&功能                           | 备注                                                                                                                 |
|-------|---------------------------------|--------------------------------------------------------------------------------------------------------------------|
| 24M10 | Paddle OCR 2.9                  | 大幅提升了文本图像版面解析能力，充分发挥文心一言语言理解优势，信息抽取整体效果相比于上一个版本提升 6%，同时新增 7 个实用的 OCR 基础模型。                                         |
| 24M9  | 千帆大模型平台 3.0                     | 针对模型调用、模型开发、应用开发三个方面进行优化升级，不仅提升了用户体验，还大幅降低了企业使用大模型的成本。                                                             |
| 24M9  | 百舸 AI 异构计算平台 4.0                | 面向万卡、十万卡集群全面升级算力管理能力。通过集群设计、任务调度、并行策略、显存优化等一系列升级，整体性能相比业界平均水平提升 30%。                                               |
| 24M6  | 文心快码 2.5                        | 在知识增强、企业研发全流程赋能、企业级安全等方面实现了能力提升。                                                                                   |
| 24M6  | 飞桨框架 3.0                        | 向下适配异构多芯，向上一体化支撑大模型的训练、推理，同时具有动静统一自动并行、编译器自动优化、大模型训推一体、大模型多硬件适配四项能力。                                               |
| 24M6  | 文心大模型 4.0 Turbo                 | 在基础大模型的基础上，进一步创新智能体技术，包括理解、规划、反思和进化，能够做到可靠执行、自我进化。上下文输入长度升级到了 128K tokens，能够同时阅读 100 个文件或网址，AI 生图分辨率提升至 1024*1024。 |
| 24M4  | ModelBuilder                    | 可以根据开发者的需求定制任意尺寸的模型，并根据细分场景对模型进行进一步精调 SFT，达到更好的效果。ModelBuilder 预置了最全面最丰富的大模型，也支持国内外第三方主流模型，是国内拥有大模型数量最多的开发平台。      |
| 24M4  | AppBuilder                      | 是目前最好用的 AI 原生应用开发工具，提前封装和预置了开发 AI 原生应用所需的各种组件和框架，大幅降低开发门槛。                                                         |
| 24M4  | 智能体开发工具 AgentBuilder            | 开发者和商家可以利用 AgentBuilder 批量生成，应用在各种各样的场景的智能体。                                                                       |
| 24M4  | 文心大模型 4.0 工具版                   | 可以体验代码解释器功能，通过自然语言交互实现对复杂数据和文件的处理与分析，还可以生成图表或文件、快速洞察数据。                                                            |
| 24M3  | 轻量级模型 ERNIE Speed、Lite、Tiny     | 轻量级大模型显著减少了参数量，更加便于客户针对特定应用场景进行模型精调。有助于客户更容易地实现预期的使用效果，也节约了大量的成本开销。                                                |
| 24M3  | ERNIE Character&ERNIE Functions | 分别适配角色扮演类应用场景（如游戏 NPC、客服对话等）、工具调用场景（对话中使用外部工具、调用业务函数等）。                                                            |
| 23M12 | 飞桨开源框架 2.6 版本和大模型重构的开发工具链       | 包括面向前后端开发，AI 应用开发以及所有开发者的新工具，打造更智能、高效、低门槛的 AI 原生应用开发新范式。                                                           |
| 23M11 | 文心一言专业版                         | 基于 4.0 的专业版具备更强的模型能力和图片生成能力，支持各种插件，适合需要使用文心一言进行代码编程、文案撰写、绘画设计等专业工作需求的用户。                                           |
| 23M10 | 文心大模型 4.0                       | 文心大模型 4.0 在理解、生成、逻辑和记忆四大能力上，都有明显提升，综合水平与 GPT-4 相比已经毫不逊色。                                                           |
| 23M10 | 百度 GBI                          | 具有支持自然语言交互、跨数据库分析和专业知识学习三方面能力，将商业分析师十几天才能完成的数据分析工作缩短到分钟级。                                                          |
| 23M9  | “灵境”插件平台                        | 实现全流程自动化，大大降低了大模型插件开发的成本。                                                                                          |
| 23M9  | “千帆大模型”平台 2.0                   | 支持大模型和数据集数量最多、工具链最完善、算力效能最佳和企业级安全四大亮点。                                                                             |
| 23M8  | 智能助理“云一朵”                       | 基于文心大模型，可快搜文件、总结 / 创作内容。                                                                                           |
| 23M6  | “擎舵”营销平台                        | 通过多模态内容制作，赋能创意生产力提升，可轻松实现文案创作、图片创作和数字人视频制作三大创意生产能力。                                                                |
| 23M5  | 文心大模型 3.5                       | 实现了基础模型升级、精调技术创新、知识点增强、逻辑推理增强等，新版本在效果、功能、性能全面提升。                                                                   |
| 23M5  | 文心一言                            | 首个可以对标 ChatGPT 的产品，实现了中文语言大模型 AI 生成式产品从无到有的突破。                                                                     |

资料来源：百度官方公众号，百度 AI 公众号，同花顺财经，新浪网，腾讯网，光大证券研究所整理

## 2.4 快手：可灵模型专注文生视频领域居业界领先

快手旗下文生视频生成模型可灵始终处于全球业界领先水平，最新基座版本更新后，带来显著画面表现力提升，并获得专家评测榜单好评。

1) 可灵在上线半年多的时间保持积极的前沿探索和模型更新，维持全球视频生成领域领先水平。可灵 24 年 6 月正式发布并上线，作为全球首个可公开体验的

真实影像级视频生成大模型，截至 25 年 1 月，可灵已完成数十次功能与效果的升级迭代，同时陆续推出多项丰富且实用的控制与编辑功能。

2) 可灵 1.6 视频生成模型文本响应度、画面美感及运动合理性，均有明显提升，图生视频更新效果提升明显。24 年 12 月，可灵 AI 宣布基座模型再升级，同时支持标准和高品质模式，特别是 1.6 模型的图生视频，内部评测比 1.5 模型整体效果提升 195%，模型物理规律真实感、语义理解得到提升，人物运动表演更强。

图 21：可灵 1.6 图生视频稳定性提升



资料来源：数字生命卡兹克公众号，可灵 AI

图 22：可灵 1.6 图生视频人物运动表演加强



资料来源：数字生命卡兹克公众号，可灵 AI

3) 可灵文生视频模型在 AGI-Eval 榜单评测中居前列，超过 OpenAI 的 Sora 模型，得分接近国产 Pixverse-V3 模型。AGI-Eval 通过构建上百条评测数据和专家级人工评测团队，对 Sora 、及国产头部视频生成模型进行专业评测。可灵 1.5 的文生视频模型在 24 年 12 月的最新测评中拿到 0.573 分，拿到第二名，高于 OpenAI 的 Sora-720p 和 Sora-1080p，仅略低于 Pixverse-V3 的 0.5732 分。

具体评价上看，与国内头部大模型（国内前三）相比，Sora 在视频-文本一致性维度、视频质量上均有小幅落后。Sora 在运动质量维度表现略好于可灵 1.6，即生成的视频画面在动态过程中的主体一致性和动态幅度更自然。在视频-文本一致性维度上，Sora 存在文本理解有误、指令遵循不符的问题，即生成的视频内容与提示词的描述不符的现象。

表 6: AGI-Eval 评测社区 24 年 12 月多模态模型（文生视频模型）榜单排行

| 排名 | 模型            | 厂商           | 开源/闭源 | 最新评测时间 | 主观得分   |
|----|---------------|--------------|-------|--------|--------|
| 1  | Pixverse-V3   | 爱诗科技         | 闭源    | Dec-24 | 0.5732 |
| 2  | Kling1.5      | 快手           | 闭源    | Dec-24 | 0.573  |
| 3  | Video-01      | Minimax      | 闭源    | Dec-24 | 0.5642 |
| 4  | Sora-720p     | OpenAI       | 闭源    | Dec-24 | 0.561  |
| 5  | Sora-1080p    | OpenAI       | 闭源    | Dec-24 | 0.548  |
| 6  | Kling1.6      | 快手           | 闭源    | Dec-24 | 0.5433 |
| 7  | Pika1.5       | Pika         | 闭源    | Dec-24 | 0.4982 |
| 8  | Vidu          | 生数科技         | 闭源    | Dec-24 | 0.4965 |
| 9  | Dream_Machine | Luma         | 闭源    | Dec-24 | 0.4942 |
| 10 | Gen3          | Runway       | 闭源    | Dec-24 | 0.4775 |
| 11 | qingying      | 智谱清言         | 闭源    | Dec-24 | 0.4592 |
| 12 | Dreamina1.2   | 字节跳动         | 闭源    | Dec-24 | 0.4562 |
| 13 | Veo2          | Google       | 闭源    | Dec-24 | 0.4376 |
| 14 | SVD           | Stability AI | 开源    | Dec-24 | 0.3842 |

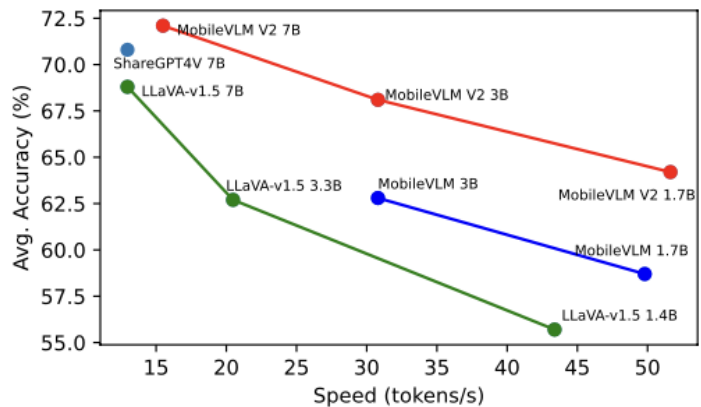
资料来源：AGI-Eval 评测社区，光大证券研究所

## 2.5 美团： MobileVLM V2 模型在移动设备环境具备优势，或仍在推进自研模型

美团联合学术界推出 MobileVLM V2 模型，有望在保持较小规模同时达到优质性能，在移动设备等环境展现优势。24 年 2 月，美团公司联合浙江大学和大连理工的研究人员共同开发的视觉语言模型 MobileVLM V2 发布。这一模型系列在保持较小模型规模的同时，达到了与更大规模 VLMs 相媲美甚至更优的性能，尤其在移动设备等资源受限的环境中展现出其独特的优势。

- 1) MobileVLM V2 在前代 MobileVLM 的基础上，通过精心设计的架构优化、改进的训练策略，以及对高质量数据集的细致策划，实现显著的性能提升。
- 2) MobileVLM 具备在移动端运行推理的能力。其高通骁龙 888 CPU 和英伟达 Jetson Orin GPU 上测量推理速度，分别获得 21.5 个 Token 和 65.3 个 Token 每秒的 state-of-the-art 性能。

图 23: MobileVLM V2 在速度和准确性上均有提升



资料来源：ADFeed, MobileVLM

美团自研大模型仍未公布名称等细节,内部或仍在推进大模型研发及摸索业务结合方向。23 年 11 月,在国内第二批通过备案的 11 家公司大模型里包含美团,但美团并未公布其大模型的名称、定位及应用案例等,整体美团在大模型的研发和投入比较隐秘。根据钛媒体消息,23 年美团已在扩张算法团队,并启动筹划单独的“平台部门”,帮助美团大模型通过具体的商业化形式落地。我们认为,美团或仍在探索大模型如何更紧密得和自身业务相结合,建议关注后续 AI 对美团内部工作提效的应用,及对美团更多业务场景的渗透。

## 2.6 哔哩哔哩：Index 自研模型在角色扮演、长文本等方面表现不俗

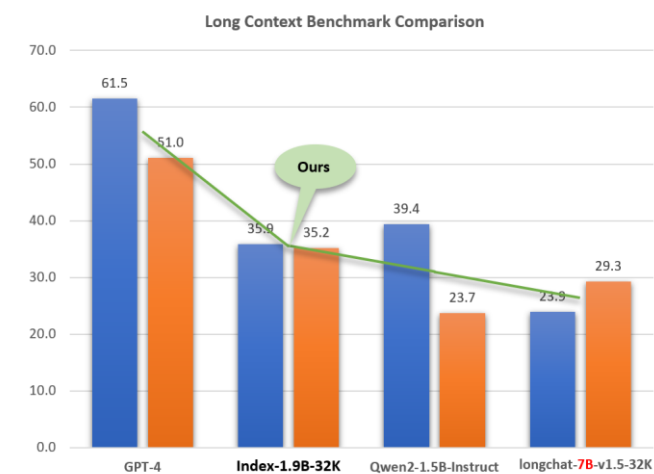
哔哩哔哩推出自研 Index 模型,在对话交互、角色扮演等方面展现出较为出色的性能。

1) Index 系列包含聊天、角色扮演等多个模型,向轻量级探索的同时,覆盖方向较为全面。24 年 6 月,哔哩哔哩发布 Index 系列模型中的轻量版本:Index-1.9B 系列,其中包含基座模型:多个评测基准上与同级别模型比处于领先;基座模型对照组;增强聊天趣味性的 Index-1.9B chat;实现 fewshots 角色扮演定制的 Index-1.9B character。24 年 9 月,哔哩哔哩开源长文本处理能力出色的 Index-1.9B-32K。

2) 哔哩哔哩 Index-70B 模型在角色扮演能力方面具备出色能力,符合哔哩哔哩自身潜在应用场景。根据 24 年 11 月在中文场景角色扮演评测集 benchmark CharacterEval 上的测试,Index-70B 角色扮演模型在该 benchmark 中均分第一,且在知识幻觉性、对话流利度、表现多样性 12 个细分维度中的 7 项中排名第一,优于情感陪伴赛道同类产品。对于 B 站来说,角色扮演模型在娱乐、教育、视频创作等方面都拥有着丰富的应用场景。

3) 哔哩哔哩 Index 模型已经应用于自身 AI 字幕等场景,期待后续模型在对 AI 接受度更高的年轻人社区得到更广泛的应用。24 年 9 月哔哩哔哩 CEO 陈睿表示,B 站将自研大语言模型 index 应用于 AI 字幕,具备中、英、韩、日、泰语等近 10 种语言的实时翻译能力,准确度接近 90%。

图 24: Index-1.9B-32K 与 GPT-4、Qwen2 等模型长文本能力对比



资料来源:哔哩哔哩技术

图 25: Index-1.9B-Character 在角色扮演林黛玉中的表现



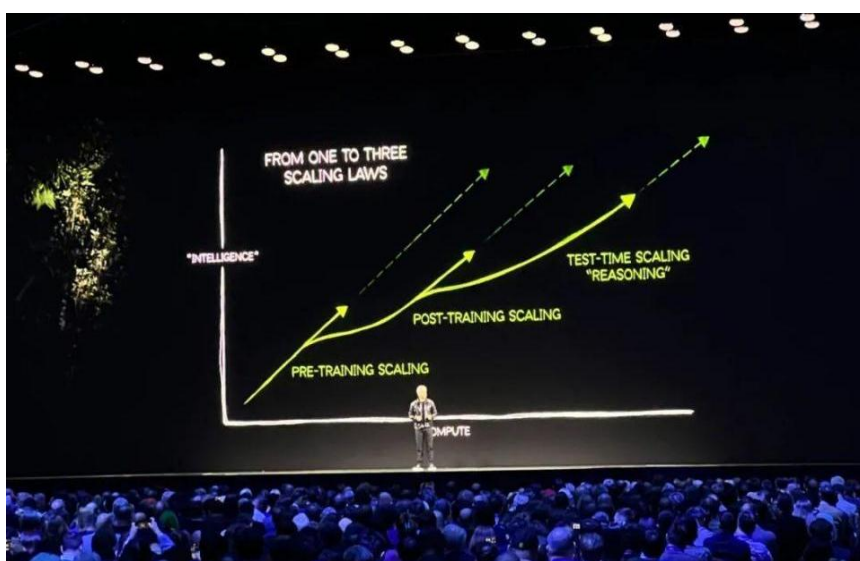
资料来源:机器之心 SOTA 模型公众号

### 3、DeepSeek “破圈” 后，对互联网大厂 AI 应用有何价值？

#### 3.1 大模型能力：R1 是起点不是终点，技术突破有望启发大厂改进模型

DeepSeek-R1 只是 DeepSeek 的第一个推理模型，其带来的突破有机会在后续研发中持续完善，在近期带来性能更强大的产品。DeepSeek-R1 证明仅用强化学习就可以在后训练阶段提升模型推理能力，后续通过在此阶段增加算力，有望满足后训练阶段的强化学习 Scaling Law(RL Scaling Law)，进而提升大模型的效率。

图 26：三种 Scaling Law（预训练、后训练和在线推理）示意



资料来源：腾讯网，CES2025

DeepSeek-R1 的研究过程已经证明对阿里 Qwen2.5 进行微调，能够提升模型能力，DeepSeek-R1 的模式有助于激发现有模型潜力。蒸馏后的 DeepSeek-R1-Distill-Qwen-32B 模型各评测指标尽管不如 DeepSeek-R1 但好于参数量级高很多的 DeepSeek-R1-Zero，也好于通义千问自己的推理模型 QwQ-32B-Preview。蒸馏后的较小模型有较低的运行成本，有利于推理模型的推广。

图 27：DeepSeek 研究过程中对 Qwen、Llama 蒸馏模型与非蒸馏模型的基准比较

| Model                         | AIME 2024   |         | MATH-500    | GPQA Diamond | LiveCode Bench | CodeForces  |
|-------------------------------|-------------|---------|-------------|--------------|----------------|-------------|
|                               | pass@1      | cons@64 | pass@1      | pass@1       | pass@1         | rating      |
| GPT-4o-0513                   | 9.3         | 13.4    | 74.6        | 49.9         | 32.9           | 759         |
| Claude-3.5-Sonnet-1022        | 16.0        | 26.7    | 78.3        | 65.0         | 38.9           | 717         |
| OpenAI-o1-mini                | 63.6        | 80.0    | 90.0        | 60.0         | 53.8           | 1820        |
| QwQ-32B-Preview               | 50.0        | 60.0    | 90.6        | 54.5         | 41.9           | 1316        |
| DeepSeek-R1-Distill-Qwen-1.5B | 28.9        | 52.7    | 83.9        | 33.8         | 16.9           | 954         |
| DeepSeek-R1-Distill-Qwen-7B   | 55.5        | 83.3    | 92.8        | 49.1         | 37.6           | 1189        |
| DeepSeek-R1-Distill-Qwen-14B  | 69.7        | 80.0    | 93.9        | 59.1         | 53.1           | 1481        |
| DeepSeek-R1-Distill-Qwen-32B  | 72.6        | 83.3    | 94.3        | 62.1         | 57.2           | 1691        |
| DeepSeek-R1-Distill-Llama-8B  | 50.4        | 80.0    | 89.1        | 49.0         | 39.6           | 1205        |
| DeepSeek-R1-Distill-Llama-70B | 70.0        | 86.7    | 94.5        | 65.2         | 57.5           | 1633        |
| <b>DeepSeek-R1-Zero</b>       | <b>71.0</b> |         | <b>95.9</b> | <b>73.3</b>  | <b>50.0</b>    | <b>1444</b> |
| <b>DeepSeek-R1</b>            | <b>79.8</b> |         | <b>97.3</b> | <b>71.5</b>  | <b>65.9</b>    | <b>2029</b> |

资料来源：Ahead of AI 博客

DeepSeek 的成功可能会加剧 AI 大模型的前沿竞争，或有望促使各大互联网公司加大对 AI 大模型的战略投入。我们认为：1) DeepSeek-R1 的成功表明，在核心技术上取得实质性突破，能够收获业界广泛好评和用户的积极参与。腾讯、阿里等国内互联网大厂或将持续加大在 AI 算法、模型架构和训练方法上的研发投入。2) DeepSeek-R1 的成功是长期技术研发积累的结果，建议关注腾讯、阿里等国内互联网大厂短期市场表现和用户规模的同时，也重点关注其长期技术积累和战略布局。

DeepSeek 的多项创新性技术突破路径已开源，其或可被复刻至各大互联网公司旗下 AI 模型中，带来模型能力的提升。正如 2.1 节中所叙述，阿里 QwQ-32B-Preview 存在的语言切换问题已被 DeepSeek-R1 部分解决，在奖励函数设计中，重点对语言一致性进行的要求，DeepSeek-R1 的模型训练思路有望成为 QwQ-32B 的借鉴。DeepSeek-R1 在后训练阶段使用 RL 的方法和 GRPO 算法、DeepSeekMoE 架构、MLA 机制、FP8 精度、MTP 方法等均可为行业所参考并用于提升自身模型能力。DeepSeek 驱动模型能力提升，有望进一步加强互联网大厂模型对内业务赋能、对外商业化输出的能力。

### 3.2 云服务：大厂已广泛支持 DeepSeek 模型，模型加速迭代有望提升云需求

DeepSeek 的成功证明海内外市场对高性能 AI 推理的强烈需求。

- 1) DeepSeek 服务器不足以长期满足大用户量使用，云厂商具备计算资源可供使用。由于 DeepSeek-R1 新模型发布后，用户访问量激增，DeepSeek 服务器无法满足大量用户的并发需求，25 年 2 月 6 日起 DeepSeek 已暂停 API 服务充值。
- 2) DeepSeek 采用宽泛、自由的 MIT 开源许可证，其一方面允许商用，使得云厂商可较为便捷地将其上线提供服务，另一方面允许修改调整模型，开源模型有望在后续得到快速迭代，进而得到可观的进步，为后续云厂商上线更多优质开源模型、提供多种模型供用户选择提供可能性。
- 3) DeepSeek-R1 受到广泛好评后，已经有多家互联网大厂接入 DeepSeek，开

始在自己的云平台提供 DeepSeek-R1 调用或部署服务。2 月 2 日以来,腾讯云、阿里云、百度智能云等互联网云服务纷纷支持调用 DeepSeek-R1 模型,部分厂商提供价格优惠,如百度智能云千帆平台推出超低价格方案,调用 R1 输入 2 元/百万 token、输出 8 元/百万 token,还可享受限时免费服务。

**表 7: 主要互联网大厂提供 DeepSeek-R1 调用服务价格**

|      | 日期      | 主要内容                                                                                            | 价格                                                                                    |
|------|---------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 腾讯云  | 2 月 2 日 | DeepSeek-R1 大模型一键部署至腾讯云「HAI」上,开发者仅需 3 分钟就能接入调用。                                                 | 调用 R1: 3.6 元/小时; 对应算力 15+TFlops SP。                                                   |
| 阿里云  | 2 月 3 日 | 阿里云 PAI Model Gallery 支持云上一键部署 DeepSeek-V3、DeepSeek-R1。                                         | 调用 R1: 输入 0.004 元/千 token; 输出 0.016 元/千 token; 免费额度: 100 万 Token (有效期: 百炼开通后 180 天内)。 |
| 百度云  | 2 月 3 日 | 百度智能云千帆平台已正式上架 DeepSeek-R1 和 DeepSeek-V3 模型,推出了超低价格方案,还可享受限时免费服务,登录百度智能云千帆 ModelBuilder 即可快速体验。 | 调用 R1: 输入 2 元/百万 token; 输出 8 元/百万 token; 限时限额免费 2 周 (至 2 月 18 日 24: 00)。              |
| 火山引擎 | 2 月 4 日 | 火山引擎为通过方舟调用 DeepSeek 模型 API 的企业提供有竞争力的价格,并提供全网最高的限流,方便业务快速落地。                                   | 调用 R1: 输入 2 元/百万 token; 输出 8 元/百万 token; 限时限额优惠 2 周。                                  |

资料来源: 腾讯云、阿里云、百度智能云、火山引擎, 光大证券研究所

我们认为, 各大互联网公司:

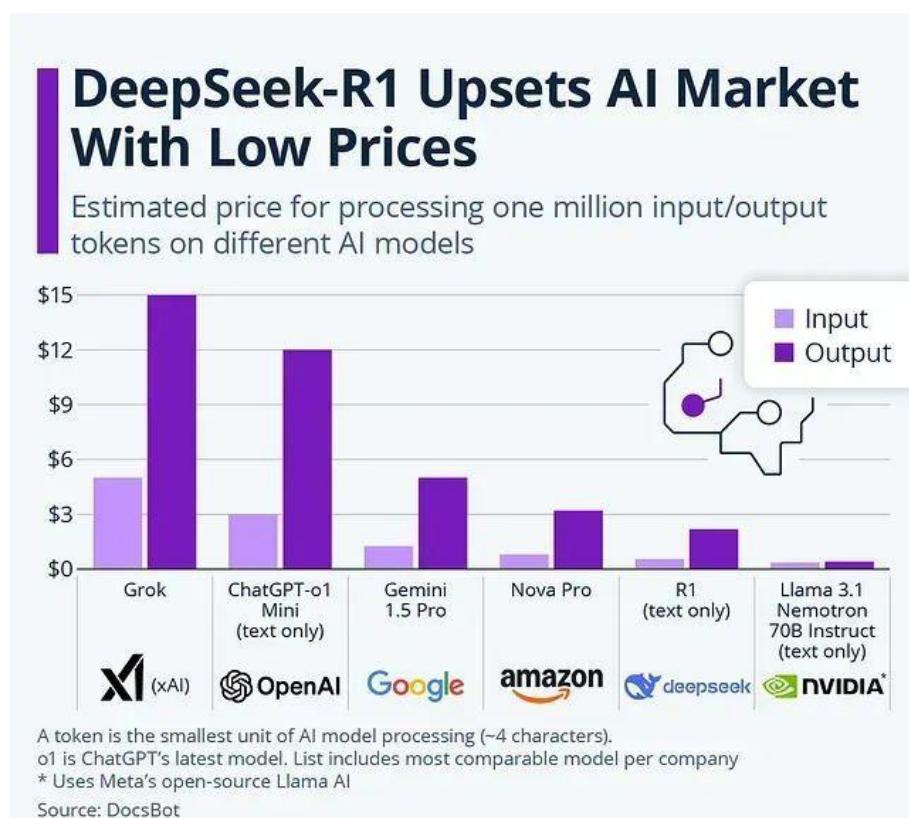
- 1) 可通过云为用户提供 DeepSeek 服务, 从而提升自身云服务付费用户量并实现云服务收入增长。
- 2) 有望伴随着 DeepSeek 的火热, 使得市场感知到并认可头部互联网公司云服务能力和未来增长空间, 有望逐步提升云业务估值水平。

具有更便捷的接入方式、更广泛的客户群体、更强的技术稳定性的云服务厂商有望率先受益。

### 3.3 AI Agent: 优质推理模型带来能力提升, 大厂 AI Agent&行业应用有望渗透

DeepSeek-R1 的推理能力和思维链能够直接赋能 AI Agent, 为其提供一定程度自主性和复杂推理能力, 助力 AI Agent 逐步完善。1) 作为优质的推理模型, DeepSeek 无需复杂提示词即可理解用户意图, 有望降低 AI Agent 的使用门槛, 提升人机交互的自然性。2) 由于 AI Agent 会自主规划任务, token 消耗相对不可控, DeepSeek 旗下旗舰模型在拥有业界领先性能的同时, 实现显著的成本降低, 有望助力维持 AI Agent 成本的相对可控。3) DeepSeek 旗下旗舰模型均开源, 使得拥有应用场景的企业或机构更容易应用 DeepSeek 模型开发适合自身垂类领域的 AI Agent, 进而使得 AI Agent 逐步被认知、被应用。

图 28: DeepSeek-R1 推理成本和主流模型对比



资料来源：腾讯云，statista

**OpenAI 已上线其 AI Agent，腾讯、阿里等互联网大厂已提供 AI Agent 服务。**1) ChatGPT 25M1 更新新功能“Tasks”，让 AI 具备一定执行力，可以替用户完成各种任务，如定时提醒天气、总结&创作文章、创建编程谜题等。2) 百度文心智能体平台、腾讯元器、讯飞星火智能体创作中心、通义智能体、字节扣子等面向企业用户提供了智能体创建平台，并开始在其 AI 智能助手界面中添加 AI Agent 入口。

**AI Agent 仍处于发展早期，具备一定不可预测性，且依赖基座模型表现，DeepSeek-R1 有望带来提升。**1) AI Agent 工作流程需链接多个 AI 步骤，用户难以确保 Agent 能否始终提供准确、符合上下文的响应。2) AI Agent 依赖基座模型需要具备较快的速度和较低的成本，特别是需要进行循环和自动重试时。

图 29：腾讯元器创建 AI Agent 页面



资料来源：腾讯元器

图 30：腾讯元器特色功能和分发渠道



资料来源：腾讯元器

表 8：AgentOps.ai 25 年 AI Agent 图谱（25 年 1 月）

| 类型                             | 应用效果                                                     | 上榜个数 |
|--------------------------------|----------------------------------------------------------|------|
| Productivity（生产力）              | 通过自动化任务、优化工作流程和提升效率来显著提高生产力，涵盖办公、学习、营销、财务等多领域的任务处理与效率提升。 | 75 个 |
| AI Agents Platform（AI 智能体平台）   | 支持开发、管理和部署智能体，提供工具和接口，助力企业和个人构建及集成智能体。                   | 74 个 |
| AI Agents Frameworks（AI 智能体框架） | 为创建高效、可扩展的 AI 智能体提供编程工具和库，包含任务规划、对话管理和数据处理等模块。           | 69 个 |
| Coding（编程）                     | 协助开发者完成代码编写、调试和优化，提升开发效率，减少人为错误。                         | 38 个 |
| Voice AI Agents（语音 AI 智能体）     | 借助语音识别和自然语言处理技术，与用户进行语音交互，广泛应用于智能家居、车载系统和客户服务等场景。        | 37 个 |
| Customer Service（客户服务）         | 提供即时客户支持、回答常见问题，并通过个性化推荐提高客户满意度。                         | 36 个 |
| Data Analysis（数据分析）            | 处理海量数据，快速生成洞察，帮助用户做出数据驱动的决策，应用于商业、科学和金融等领域。              | 35 个 |
| Digital Workers（数字化工作者）        | 专注执行重复性高的任务，如数据输入、文档整理和流程自动化。                            | 31 个 |
| Personal Assistant（个人助理）       | 帮助用户管理日程、发送提醒、执行简单任务，提供个性化建议，提升个人日常效率。                   | 30 个 |

资料来源：AI 信息 Gap 公众号，光大证券研究所

### 3.4 AI+广告：AIGC 重塑营销链条，更强的模型效果提升广告自动化能力

海外互联网大厂已经在广告主侧（B 端）和消费者侧（C 端）全面重塑营销链条，从而提升广告投放效率和增加广告创收。

#### 1) B 端：

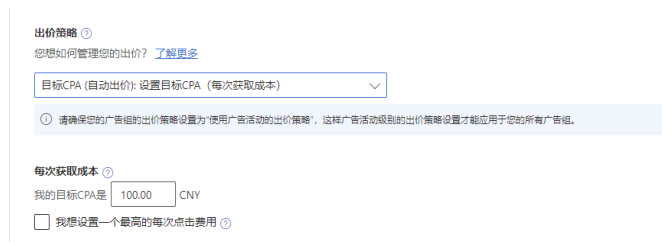
过去依赖人工经验和传统计算的营销决策，正在被具备高度数据理性的 AI 系统所取代。微软广告的动态搜索广告（DSA）系统会自动为每一个落地页动态创建广告，并基于 AI 驱动自动识别搜索与对话背后的用户意图，从而更高效地和商家广告精准匹配，最终达成获客成本的下降和投资回报率的提升。Meta 基于 AI

驱动的自动化广告产品 Advantage+已经让 Meta 的广告收入实现连续的强劲增长。

2) C 端：在 AI 生成内容中创新性地引入广告，在广告组件中的交互中运用生成式 AI。

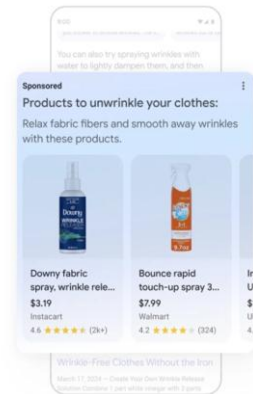
23 年谷歌就将广告引入到对话型 AI 产品中，24 年更进一步，AI 生成的搜索结果概要 (Overviews) 就成为一个重要的广告位置，其图像识别工具 Lens 中推出购物广告，目标是吸引更多电商类客户投放广告。AI 对于广告承载页的交互方面，用户在谷歌搜索装修和家具购买时，谷歌会允许用户提交一系列客厅照片，然后智能地向他们推荐适合用户目标房间的相关家具。

图 31：提出自动竞价策略示例



资料来源：微软广告

图 32：谷歌 AI 概要广告



资料来源：Marketing Global 公众号

国内互联网大厂已有 AI+广告投放方面、AI+广告素材生成方面布局，DeepSeek 系列旗舰模型在提升性能的同时全面开源，有望助力互联网大厂进一步增强自动化广告投放能力。

1) 对于百度，24 年 12 月百度商业系统升级为“百度伴飞”，基于文心大模型，整合多样化的 AI 能力，助力实现高品质品牌宣传，并带来视频广告点击率、完播率等投放效果提升。

2) 对于腾讯，腾讯广告妙思由腾讯混元大模型提供基底支持，通过其对语义的精准理解和表达，生产稳定实用的广告创意素材，降低广告优化师制作广告的成本。

3) 对于哔哩哔哩，25 年 1 月推出首个商业化 AIGC 平台“星辰 AI”，能够智能化地生成图片素材，同时优化广告创意的迭代过程，有望提高广告素材的转化率。

图 33：百度城市名片智能体示例



资料来源：首席营销官公众号

图 34：腾讯广告妙思具备文生视频等全面广告创意素材生成能力



资料来源：卫夕指北公众号，腾讯广告妙思

我们认为，广告作为高毛利业务板块，AI 赋能空间理想。腾讯等国内互联网公司此前在 AI+广告赋能领域相较于海外大厂较慢，部分因模型性能等 AI 技术能力相较海外领先水平有一定差距，且应用 AI 赋能广告的实践晚于海外大厂。

若借力 DeepSeek 等高性能模型带来的模型水平提升，有望伴随着 AI 广告的应用实践，逐步加深 AI 自动化广告的渗透，将大模型更深度地用于分析用户在自身生态上的行为，精准总结画像进而优化广告定向投放，促进广告点击率等指标增长，进而提升广告业务收入和利润水平的增长率。25 年 1 月腾讯集团年会中，董事会主席兼 CEO 马化腾表示，看好 AI 未来给广告带来的空间。

## 4、投资建议

DeepSeek-R1 因其赶超 OpenAI 推理模型 o1 的强大性能、多项算法和工程上的实质性突破、全面开源并推出便于扩散的免费 C 端产品，实现在业界和大众间的“破圈”。我们认为，复盘 DeepSeek-R1 的成功，建议关注投资主线：

- 1) DeepSeek-R1 作为中国模型首次受到海外 AI 科技界全面推崇认可。Ray Dalio《2024 年大国指数》显示，科技创新一项中美国（高于平均 1.9 标准差）仅略超中国（1.8），是中美差距最小的项之一。关注 DeepSeek-R1 等科技成果催化下的中概资产价值重估。
- 2) 关注 DeepSeek 系列之外在技术进展等方面领先的中国大模型，推荐：阿里巴巴-W：旗下最新旗舰模型 Qwen2.5-Max 已能赶超业界领先的模型，实验性研究推理模型 QwQ-32B-Preview 在数学和编程等领域已取得显著进步。腾讯控股：上线业界混元 3D AI 创作引擎，组织架构调整后 AI 战略更清晰。快手-W：文生视频生成模型可灵始终处于全球业界领先水平。百度集团-SW：文心大模型多次迭代，期待 25 年新版本面世。
- 3) 关注通过云为用户提供 DeepSeek 服务实现收入增长、估值提振的头部互联网公司，推荐：阿里巴巴-W：具有更便捷的接入方式、更广泛的客户群体、更强的技术稳定性的云服务厂商有望率先受益。
- 4) 关注借力 DeepSeek 增强 AI 模型能力，提升 AI+广告渗透，进而获得广告增长的头部互联网公司，推荐：腾讯控股、快手-W、哔哩哔哩、百度集团-SW：广告作为高毛利业务板块，AI 赋能空间理想。

## 5、风险提示

### 1、AI 技术研发和产品迭代不及预期：

若后续 DeepSeek 及国内阿里通义等其他大模型研发进展受阻，进而影响产品开发迭代进度，或影响行业长期发展预期。

### 2、AI 行业竞争加剧风险：

伴随 DeepSeek 将优质模型开源，大量公司和机构有机会借力提升自身模型能力，或带来 AI 大模型领域竞争加剧。

### 3、商业化进展不及预期风险：

尽管 DeepSeek 带来模型性能的提升，但能否以及如何将模型能力进行进一步的商业化变现，尚具备一定不确定性。

### 4、国内外政策风险：

尽管 DeepSeek 带来模型性能的提升，整体 AI 大模型仍可能具备生成内容的不可靠性风险、训练数据侵权风险，带来政策监管风险。

## 行业及公司评级体系

|         | 评级  | 说明                                                       |
|---------|-----|----------------------------------------------------------|
| 行业及公司评级 | 买入  | 未来 6-12 个月的投资收益率领先市场基准指数 15%以上                           |
|         | 增持  | 未来 6-12 个月的投资收益率领先市场基准指数 5%至 15%；                        |
|         | 中性  | 未来 6-12 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；                   |
|         | 减持  | 未来 6-12 个月的投资收益率落后市场基准指数 5%至 15%；                        |
|         | 卖出  | 未来 6-12 个月的投资收益率落后市场基准指数 15%以上；                          |
|         | 无评级 | 因无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使无法给出明确的投资评级。   |
| 基准指数说明： |     | A 股市场基准为沪深 300 指数；香港市场基准为恒生指数；美国市场基准为纳斯达克综合指数或标普 500 指数。 |

## 分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

## 分析师声明

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，并对本报告的内容和观点负责。负责准备以及撰写本报告的所有研究人员在此保证，本研究报告中任何关于发行商或证券所发表的观点均如实反映研究人员的个人观点。研究人员获取报酬的评判因素包括研究的质量和准确性、客户反馈、竞争性因素以及光大证券股份有限公司的整体收益。所有研究人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中的具体推荐意见或观点有直接或间接的联系。

## 法律主体声明

本报告由光大证券股份有限公司制作，光大证券股份有限公司具有中国证监会许可的证券投资咨询业务资格，负责本报告在中华人民共和国境内（仅为本报告目的，不包括港澳台）的分销。本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格编号已披露在报告首页。

中国光大证券国际有限公司和 Everbright Securities(UK) Company Limited 是光大证券股份有限公司的关联机构。

## 特别声明

光大证券股份有限公司（以下简称“本公司”）成立于 1996 年，是中国证监会批准的首批三家创新试点证券公司之一，也是世界 500 强企业——中国光大集团股份公司的核心金融服务平台之一。根据中国证监会核发的经营证券期货业务许可，本公司的经营范围包括证券投资咨询业务。

本公司经营范围：证券经纪；证券投资咨询；与证券交易、证券投资活动有关的财务顾问；证券承销与保荐；证券自营；为期货公司提供中间介绍业务；证券投资基金代销；融资融券业务；中国证监会批准的其他业务。此外，本公司还通过全资或控股子公司开展资产管理、直接投资、期货、基金管理以及香港证券业务。

本报告由光大证券股份有限公司研究所（以下简称“光大证券研究所”）编写，以合法获得的我们相信为可靠、准确、完整的信息为基础，但不保证我们所获得的原始信息以及报告所载信息之准确性和完整性。光大证券研究所可能将不时补充、修订或更新有关信息，但不保证及时发布该等更新。

本报告中的资料、意见、预测均反映报告初次发布时光大证券研究所的判断，可能需随时进行调整且不予通知。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。客户应自主作出投资决策并自行承担投资风险。本报告中的信息或所表述的意见并未考虑到个别投资者的具体投资目的、财务状况以及特定需求。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，本公司及作者均不承担任何法律责任。

不同时期，本公司可能会撰写并发布与本报告所载信息、建议及预测不一致的报告。本公司的销售人员、交易人员和其他专业人员可能会向客户提供与本报告中所载观点不同的口头或书面评论或交易策略。本公司的资产管理子公司、自营部门以及其他投资业务板块可能会独立做出与本报告的意见或建议不相一致的投资决策。本公司提醒投资者注意并理解投资证券及投资产品存在的风险，在做出投资决策前，建议投资者务必向专业人士咨询并谨慎抉择。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问或金融产品等相关服务。投资者应当充分考虑本公司及本公司附属机构就报告内容可能存在的利益冲突，勿将本报告作为投资决策的唯一信赖依据。

本报告根据中华人民共和国法律在中华人民共和国境内分发，仅向特定客户传送。本报告的版权仅归本公司所有，未经书面许可，任何机构和个人不得以任何形式、任何目的进行翻版、复制、转载、刊登、发表、篡改或引用。如因侵权行为给本公司造成任何直接或间接的损失，本公司保留追究一切法律责任的权利。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

光大证券股份有限公司版权所有。保留一切权利。

## 光大证券研究所

### 上海

静安区新闻路 1508 号  
静安国际广场 3 楼

### 北京

西城区武定侯街 2 号  
泰康国际大厦 7 层

### 深圳

福田区深南大道 6011 号  
NEO 绿景纪元大厦 A 座 17 楼

## 光大证券股份有限公司关联机构

### 香港

中国光大证券国际有限公司  
香港铜锣湾希慎道 33 号利园一期 28 楼

### 英国

Everbright Securities(UK) Company Limited  
6th Floor, 9 Appold Street, London, United Kingdom, EC2A 2AP