东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 02 期：AGI 是什么？

2026 年 07 月 28 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn

投资要点

- **AGI 的核心研究问题，是如何定义、衡量并判断通用人工智能何时到来。**全球机构对 AGI 的理解主要落在四类锚点：以 OpenAI、麦肯锡为代表的经济价值口径，以 DeepMind 为代表的能力分级口径，以 Chollet/ARC 为代表的技能获取效率口径，以及 LeCun、李飞飞等更强调世界理解与具身交互的口径。不同定义并非简单对错，而是回答“智能以结果、能力、学习效率还是世界模型为准”的不同问题。

- **从机构判断看，AGI 并非一个单一时间点，而更像能力阶梯。**OpenAI 五级路线图以经济价值和自主程度为主轴，从对话者、推理者、智能体、创新者走向组织级 AI；DeepMind《Levels of AGI》则以性能和通用性二维刻画，从 L0 到 L5 衡量能力层级。前者直观、便于映射商业化进度，后者更审慎、便于学术比较。

- **围绕通往 AGI 的路径，分歧集中于现有范式能否外推。**激进阵营认为规模化、测试时推理、强化学习和智能体系统可继续推进；审慎阵营提示现有模型缺世界模型、因果和稳健泛化，可能需要新范式甚至物理具身；务实阵营更重应用落地，对时间表保持克制。

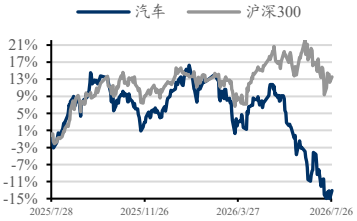
- **AGI 的衡量与经济价值仍缺统一答案。**传统基准趋于饱和，ARC-AGI、METR、GDPval、DeepMind L0-L5 等新标尺各自测量推理、任务时长、经济活动或能力层级，但行业尚缺权威评测机构。经济测算方面，从十年拉动 GDP 约 1%的保守估计，到每年数万亿美元增加值乃至全自动化后的增长爆发，差异主要来自劳动替代比例、用例边界和自动化假设不同。

- **产业映射上，**短期更可观察的是高价值职能渗透和推理、智能体、具身智能等工程化拐点；中长期需跟踪测试时计算、强化学习、世界模型、机器人等路径能否跨越当前可验证任务边界。

- **我们认为,数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**

- **风险提示：**技术迭代不及预期的风险；大模型厂商 ARR 增速放缓的风险；云服务商资本开支不及预期的风险；智能驾驶及机器人商业化落地不及预期的风险。

行业走势



相关研究

《AI 系列深度 01 期：从数字 AI 通用底座到物理 AI 真实世界闭环》

2026-07-27

《整车主线周报：本周 SW 商用载客车表现较好，重卡 6 月海关出口数据公布》

2026-07-27

内容目录

图表目录

1. AGI 定义：四类锚点与底层分歧

通用人工智能 AGI 通常被理解为“具备跨领域通用能力、能像人一样学习并解决新问题的系统”。但“通用”“人类水平”至今没有统一的可量化判定标准。本质分歧并非定义本身，而是不同机构正在用完全不同的评价体系衡量 AGI 的发展进度。

全球主流“理解”大致落在四类锚点上：其一“经济价值”，关注 AI 能否替代多数有经济价值的工作、不追问内部机制（如 OpenAI、麦肯锡）；其二“能力分级”，从性能与通用性两维度打分（如 DeepMind 的 L0–L5 分级体系）；其三“技能获取效率”，强调“在陌生任务上快速学会”而非堆叠已有技能（如 Chollet 的 ARC 测试）；其四“对世界的理解”，认为没有世界模型、缺乏空间与因果的系统不算真正通用（以 LeCun、李飞飞等学者为代表）。

越靠近“结果 / 经济价值”，越容易宣布“接近 AGI”；越靠近“机制 / 世界理解”，门槛越高、越难达标。这正是同一进展可被乐观者称为“前夜”、被审慎者判为“尚在山脚”的根本原因。

表1：全球主要机构对 AGI 的理解谱系

机构 / 代表	类别	对 AGI 的独特理解 / 提法	锚点
OpenAI，章程，	前沿实验室	在多数有经济价值工作上超越人类的高度自主系统	经济价值
DeepMind，Morris 等，	实验室/学术	按性能×通用性分级的能力本体，L0–L5，	能力分级
Anthropic Amodei	前沿实验室	“强大 AI”：多领域强于诺奖得主、具全部虚拟工作接口	能力上限
Meta / LeCun	实验室/学术	反对 AGI 提法，须建立在世界模型之上的人类水平 AI	世界理解
xAI / 马斯克	前沿实验室	“比最聪明的人更聪明”即接近 AGI	超越个体
SSI / Ilya	前沿实验室	直达“安全超级智能”；规模化时代结束、研究时代开启	研究范式
微软 / Suleyman	前沿实验室	“人本超级智能”：服务人类、不可完全自主与自我改进	可控性
Chollet / ARC	学术/标准	智能 = 在未知任务上的技能获取效率	泛化效率
Legg & Hutter	学术	机器智能的形式化定义“universal intelligence”	形式化定义
张钹，清华/中科院院士，	学术	第三代人工智能，知识+数据双驱动，可解释、安全性高	可解释
李飞飞，斯坦福/World Labs，	学术/创业	“空间智能”：语言之外、须世界模型补齐空间与因果	空间智能
上海人工智能实验室/周伯文	中国科研	“通专融合”，ANI→ABI→AGI，AI-45° 平衡律	通专融合
阿里 / 吴泳铭	中国企业	AGI 只是起点、ASI 才是终极；三阶段演进	起点论/ASI
字节 / Seed Edge	中国企业	AGI 长期研究五方向，含超越 Next-Token 的新范式，	新范式探索
腾讯 / 马化腾	中国企业	务实，不追刷分；把 AI 嵌入场景、让场景产生价值	应用务实
百度 / 李彦宏	中国企业	怀疑“AGI 是否存在”，无模型包打天下，重应用	应用/怀疑
智谱 / 张鹏	中国企业	AGI 路径 L1 预训练→L2 对齐推理→L3 工具/Agent	路径分层
MiniMax / 闫俊杰	中国企业	“Intelligence with everyone”，智能普惠、与用户共创	智能普惠
月之暗面 / 杨植麟	中国企业	以 AGI 为主营，长上下文 + 智能体集群	长上下文
DeepSeek / 梁文锋	中国企业	“模型即智能体”，坚持开源探索 AGI	开源/效率
斯坦福 HAI《AI Index》	学术/标准	以年度可测量指标刻画 AI 进展与影响	可测量
Brynjolfsson，斯坦福，	经济学	以生产力与经济价值衡量 AI 的现实影响	经济价值

数据来源：各机构与人物公开资料、主流 AI 深度访谈，东吴证券研究所

2. 全球机构判断：AGI 是一条能力阶梯而非单一时点

全球主要 AI 公司、实验室与科研机构对 AGI 的判断，可按其公开立场归为四类阵营：激进（近端）、审慎（远端）、务实（应用）、具身（物理）。

表2：全球主要 AI 公司与科研机构对 AGI 的判断一览

机构 / 人物	类别	对 AGI 的核心判断	时间表 / 阵营
OpenAI / Altman	前沿实验室	已知如何构建 AGI	≤2029 · 激进
DeepMind / Hassabis	前沿实验室	当前智能体只是“练习”	约 2030 · 激进
Anthropic / Amodei	前沿实验室	“强大 AI”，强于诺奖得主	2026–2027 · 激进
Meta / LeCun	实验室/学术	LLM 是岔路，须世界模型	不设时点 · 审慎/具身
xAI / 马斯克	前沿实验室	超越最聪明人类即 AGI	2026–2029 · 激进
SSI / Ilya	前沿实验室	规模化时代结束，转研究	研究驱动 · 审慎
微软 / Suleyman	前沿实验室	反对 AGI 竞赛、重可控	治理 · 务实
Hinton	学术界	AGI 临近但风险巨大	约 2028–2030 · 审慎
吴恩达	学术界	AGI 被过度炒作	数十年 · 审慎
李飞飞 / Chollet	学术界	缺世界模型 / 刷分≠泛化	强调机制 · 审慎
上海 AI Lab / 周伯文	中国科研	通专融合通向 AGI	务实
阿里 / 吴泳铭	中国企业	AGI 是起点、ASI 终极	投入激进 · 表态务实
字节 / Seed Edge	中国企业	长期研究五方向	务实/探索
腾讯 / 马化腾	中国企业	务实、场景价值驱动	务实
百度 / 李彦宏	中国企业	怀疑“AGI 存在”	务实/怀疑
智谱 / 张鹏	中国企业	L1→L2→L3 路径	务实
MiniMax / 闫俊杰	中国企业	智能普惠、与用户共创	务实
DeepSeek / 梁文锋	中国企业	模型即智能体、开源	务实/效率
宇树 / 具身阵营	中国企业	智能须在物理交互中习得	具身
NVIDIA / 黄仁勋	产业	由“五年内”改口“已实现”	定义漂移
斯坦福 HAI / VC	产业/投资	把 AGI 距离映射为估值	落地导向

数据来源：各机构与人物公开发言、主流 AI 深度访谈播客，东吴证券研究所

3. 通往 AGI：现有范式能否外推仍未收敛

3.1. 各机构已公开的分级框架

已公开的 AGI 相关分级框架，绝大多数是“能力 / 安全 / 落地成熟度”的等级阶梯，回答的是“能力到了什么程度”；目前业内尚未对“用什么底层技术、按什么顺序实现 AGI”的完整技术进化路径形成共识。

表3：全球主要机构已公开的 AGI 相关分级 / 路线图盘点

机构	框架名称	内容要点	性质
OpenAI	五级路线图 2024	对话→推理者→智能体→创新者→组织，L5=AGI，	能力阶梯
Google DeepMind	《Levels of AGI》 2023	L0–L5，按性能×通用性分级	能力阶梯
阿里 / 吴泳铭	三阶段，2025 云栖，	智能涌现→自主行动→自我迭代，ASI 为终极，	能力阶梯
智谱 / 张鹏	L1–L3	预训练→对齐推理→工具/Agent	能力阶梯
Anthropic	AI 安全等级 ASL	ASL-1…ASL-4+，随能力增强而加严管控	安全分级
Salesforce	智能体成熟度模型	4 级，面向企业落地的 Agent 成熟度	落地成熟度
北京通研院 / 朱松纯	AGI 标准·评级·测试·架构	一套 AGI 评级与测试体系	学术标准
微软、Meta、百度、华为	—	均未发布官方 AGI 分级路线图	未发布
腾讯、商汤、科大讯飞、字节	—	重模型迭代与应用，未发布官方分级	未发布
NVIDIA	—	提“Physical AI 下一浪”，无正式 AGI 分级	未发布

数据来源：各机构官方发布与公开资料，东吴证券研究所

OpenAI 于 2024 年中向员工内部公布、并经媒体披露的“通往 AGI 五级框架”，是目前传播最广的 AGI 分级方案之一。其延续 OpenAI 章程对 AGI 的经济学定义——“在多数有经济价值的工作上超越人类的高度自主系统”，以“AI 能在多大程度上、多自主地承担有经济价值的工作”为主轴，把通往 AGI 的进程切成五级：Level 1“对话者”——具备自然语言对话能力，即当前以 ChatGPT 为代表的形态；Level 2“推理者”——能在不借助外部工具的情况下完成博士水平的多步推理与问题求解，以 o 系列推理模型为标志；Level 3“智能体”——能代表用户自主采取行动、连续工作数天、调用工具完成复杂任务；Level 4“创新者”——能辅助乃至独立做出发明创造、推动科学发现；Level 5“组织”——能独立完成一整个组织的全部工作，被视为 AGI 的最终形态。

其分级思路的核心，是把“智能”还原为“可被替代的经济活动”——从“会聊天”到“会推理”，再到“会自主行动”“会创新”，最终到“能运转一个组织”，每升一级，AI 承担工作的复杂度、自主性与时间跨度都显著跃升。这一框架的优点是直观、可与商业化进度一一对应，OpenAI 多次据此自我定位，称当前正处于 L2 向 L3 过渡；其局限在于完全以“结果/经济产出”为锚点、不回答“智能从何而来”，因而也最容易让人得出“已接近 AGI”的乐观判断——Altman 据此称“已确信知道如何构建 AGI”。

DeepMind 团队于 2023 年 11 月在论文《Levels of AGI: Operationalizing Progress on the Path to AGI》中提出了一套更学术、更强调“可操作度量”的分级体系。它最大的特点是用两个维度而非单一标尺刻画 AGI：一是性能，AI 在任务上达到的能力深度，二

是通用性，AI 能力覆盖任务的广度。在性能维度上设六个等级——Level 0 无 AI、Level 1 新兴，等于或略优于无技能人类、Level 2 胜任，达到熟练成年人前 50%、Level 3 专家，前 10%、Level 4 大师，前 1%、Level 5 超人，超越 100% 人类；并区分“狭义”与“通用”两列：例如 AlphaFold 属“狭义超人”，而当前的前沿大模型大致处于“通用·新兴”一格。

DeepMind 分级思路的核心，是把“是否达到 AGI”从一个模糊的是非题，转化为“在性能×通用性坐标系里走到哪一格”的可测量问题——只有当某系统在足够宽的任务范围上稳定达到较高性能等级时，才被认定为相应级别的 AGI。论文同时给出六条配套原则，其中关键两条是：聚焦“能力”而非“实现方式”，不要求与人脑机制一致、不要求有意识，以及聚焦“认知与元认知任务”、明确机器人具身“不作为达成 AGI 的前提”、物理能力“有益但非必需”。相比 OpenAI 偏结果导向、与商业进度挂钩的“经济阶梯”，DeepMind 的方案更严谨、更便于横向比较与学术讨论，但也更保守——按其标尺，当前模型仅处于“新兴 AGI”，距“胜任级通用 AGI”仍有明显距离。

3.2. 主要技术路径

通往 AGI 的技术路径尚无统一分类法。我们按照归纳的技术范畴并列各家押注方向，仅作对照，不含排名与成功概率判断。其中“智能体 Agent”不属于“智能从何而来”这一层，而是“系统 / 编排”层——把现成模型包上规划、工具、记忆以自主完成长任务；围绕它的核心争议是“AGI 是模型问题还是系统问题”。

表4：各方押注的技术路径对照

技术范畴	主张 / 代表机构	依据与说明
训练时规模化	早期 OpenAI、Scaling 信徒	扩大算力/数据/参数 + 预训练；争议：是否撞“数据墙”
测试时计算与强化学习	OpenAI，o 系列、DeepSeek-R1	让模型“想得更久”、用 RL 自提升推理；新的 Scaling 轴
新架构与学习目标	LeCun JEPA、字节 Seed Edge	换掉“预测下一词”目标与 Transformer 架构
世界模型与具身	李飞飞、LeCun、宇树	从感知 / 物理交互学到有因果、空间的世界表征
神经符号与可解释	张钹，第三代 AI、ARC 方向	神经网络 + 符号 / 程序推理，强泛化与可解释
自我改进 / 开放式进化	SSI、Clune、Recursive、AI Scientist	AI 改进 AI、自动做研究、开放式探索
智能体，系统/编排层，	几乎所有实验室 + 创业生态	通过规划、工具调用和记忆机制支持长程自主；争议：模型问题还是系统问题

数据来源：各实验室公开路线图与论文，东吴证券研究所

3.3. 三大分歧的根源

分歧的核心是“现有范式能否外推”。激进 / 近端阵营（如 OpenAI、Anthropic、xAI）认为规模化叠加推理可外推到 AGI；审慎 / 远端阵营（如 LeCun、吴恩达、Hinton、Ilya）认为现有大模型缺世界模型与因果，规模再大也难跨越“理解”鸿沟；务实阵营（以多数中国大厂为代表）重应用落地、对“何时到达”不下断言；具身阵营（如中国具身企业、LeCun）主张智能须在物理世界习得。三者并非简单对错，而是对“范式能否外推、用什么衡量、智能如何定义”给出不同回答；叠加商业激励差异，表面分歧被进一步放大。

3.4. 时间表预测的方法与口径

我们对“AGI 何时到达”采用了多种预测方式，预测来自不同人群、定义与方法。乐观者的预期约在 2030 年左右，但 AI 从业者的预期则十分保守，认为 AGI 的到来时间的中位数约在 2047 年左右。

表5：AGI 到达时间的五类预测方法论对照

预测方法	怎么做	关键假设 / 弱点	结论
专家问卷 AI Impacts/Grace	调查数千名 AI 研究者、聚合其概率分布	框架效应极大：同一问卷中，措辞差异可能显著影响结论	HLMI 中位数约 2047
预测平台 Metaculus	社区做概率预测、持续更新，算法加权取社区中位数	高度依赖 resolution 口径	“通用 AI”≈2028、“strong AGI”≈2033，2025 后移，
任务时长外推 METR	测 AI 自主完成任务的人时长度，50%可靠，	只覆盖软件/推理任务	约每 7 个月翻倍，外推 5 年内做“月级”软件任务
算力 / 生物锚 Cotra、Epoch	估变革性 AI 所需训练算力 + 预测有效算力增长	锚点假设跨度极大	给出年份的概率分布，区间很宽，
实验室 CEO 表态	公开定性判断	低门槛、含商业激励	Amodei≈2027 / Altman≤2029 / Hassabis≈2030

数据来源：AI Impacts、Metaculus、METR、Epoch、各 CEO 公开发言，东吴证券研究所

4. AGI 衡量：新基准涌现，统一标准仍缺位

4.1. 衡量 AGI 的主要困难

衡量 AGI 本身就是一个有大分歧、甚至可能没有客观答案的难题，主要有四重困难，均为业界可观察事实：其一“基准饱和”，到 2026 年初前沿模型在数学、代码、知识问答类基准上普遍超过 90%，已难区分高下；其二“Goodhart 定律”，一个指标一旦成为优化目标就不再是好指标——越刷分，真实能力未必越强；其三“数据污染”，训练数据可能已含测试题，且污染检测缺乏统一标准；其四“依赖自报”，业界常被迫依赖厂商自行公布的成绩。

4.2. 主流评测基准：测法与采纳度

这些基准基本都是大厂旗舰模型卡（model card）实际报告的“行业标准套装”：2026 年初 Claude Opus 4.6、Gemini 3 等旗舰模型卡同时报告 SWE-bench、ARC-AGI-2、GPQA、HLE、GDPval 等，其中 ARC-AGI 已被 Anthropic、Google、OpenAI、xAI 四家写进模型卡。它们普遍饱和很快：ARC-AGI-2 在成本受限的 2025 年竞赛口径下前沿成绩仅约 24%，而不受限的旗舰模型到 2026 年初已大幅跃升，Gemini 3 Deep Think 84.6%、Claude Opus 4.6 68.8%。

表6：衡量 AI / AGI 进展的主要基准对照

基准 / 创建方	怎么测，原理，	谁在用 / 采纳度
ARC-AGI Chollet/ARC Prize	看几组“输入网格→输出网格”例子推出规则、套到新题；抗背题	四大实验室写进模型卡，事实标准，
GPQA Diamond，学术，198 题，	博士级理化生多选题，“防搜索”：专家答对、非专家查网仍错	大厂高频报告
HLE 人类最后考试 CAIS+Scale AI	约 2500 题、近千名专家出题、封闭式防搜索难题	前沿模型卡常报
FrontierMath Epoch AI	数学家原创未发表难题，给 Python 环境跑代码验证、自动判分	前沿数学评测、防污染
SWE-bench Verified，普林斯顿，	真实 GitHub 问题，模型交补丁，看仓库单元测试是否通过	编码事实标准，已近饱和，
GDPval OpenAI	跨职业真实任务，专家私出题、人工评审、Elo 评分	大厂开始报告
METR 任务时长，非营利 METR，	测 AI 自主完成任务的“人时长度”，50%可靠，	研究型指标，非打榜基准

数据来源：各基准创建方与大厂模型卡，东吴证券研究所

4.3. 各方提出的替代标尺

针对衡量难题，各方提出不同的“替代标尺”：经济价值（能否真正完成有偿工作，如 GDPval）、泛化效率（ARC-AGI）、自主任务时长（METR）、能力分级（DeepMind L0–L5）、对抗 / 动态评测（每次随机化参数以抗记忆）。这些标尺尚未收敛为统一标准——“如何衡量 AGI”本身仍是开放问题，也是判断“是否到达 AGI”的根本障碍之一。

5. 经济影响：测算口径决定价值判断

“AGI 能带来多大经济价值”取决于按什么口径测算——劳动替代比例、GDP 增长、用例增加值，或“全自动化后增长爆发”。

5.1. 主要测算方法：原理与结果

① 用例增加值法。原理：自下而上，把生成式 AI 拆成约 60+ 类具体用例、再映射到企业各职能，逐一估算每类用例带来的增加值，提效、增收、降本，加总得年度经济价值。结果：生成式 AI 每年可新增 2.6–4.4 万亿美元增加值，约 75% 集中于客户运营、营销、软件工程、研发四大职能；若叠加“嵌入既有软件”的影响，潜在价值大致翻倍至约 6.1–7.9 万亿美元/年。

② 职业 / 任务暴露法。原理：不直接算钱，而是把职业拆成任务、逐一判断 AI 能否胜任，统计 “多大比例的工作任务会受到影响，即暴露”。结果：OpenAI，Eloundou 等，——约 80% 美国劳动力至少 10% 任务被影响、约 19% 的人过半任务被影响；IMF——全球约 40% 岗位暴露，发达经济体约 60%，其中约一半受益、一半承压。它回答“多少工作被触及”，不直接给市场规模。

③ 宏观 task-based 法。原理：同样基于任务，但用 Hulten 定理严格推导——总生产力提升 ≈，被自动化任务占比，×，这些任务平均成本节约，只计“确实能自动化”的那部分、且只算成本节约。结果：未来十年 AI 对全要素生产率 TFP 提升不超过约 0.66%、对 GDP 拉动约 1%，比偏乐观的自下而上估算，如麦肯锡，低约一个数量级；并指出早期样本多是“易自动化”任务，该估计可能还偏高。

④ 行业咨询综合法。原理：综合生产力提升与消费侧效应，估算 AI 到 2030 年对全球经济的总贡献。结果：到 2030 年 AI 可为全球经济贡献约 15.7 万亿美元，生产力提升约 6.6 万亿 + 消费侧约 9.1 万亿。

⑤ 全自动化→增长爆发，Davidson、Aghion-Jones 增长模型。原理：这才是真正“AGI 口径”的问法——若 AI 能替代几乎全部人类劳动，含 AI 研发本身，在标准内生增长模型里劳动不再是瓶颈、自动化与创意相互复利，增速会“爆发式”跃升。结果：不是点估计、而是情景——可能出现远高于历史 2–3% 的年增长，数十个百分点量级，但高度依赖瓶颈假设，鲍莫尔成本病、实体与监管约束、能源，争议极大。

表7：AI / AGI 经济价值的主要测算方法对照

测算方法	代表机构与结果	出处
用例增加值法	麦肯锡：每年 $2.6–4.4 万亿，叠加嵌入软件约 $6.1–7.9 万亿	McKinsey 在 2023 年
职业/任务暴露法	OpenAI：80% 劳动力≥10% 任务；IMF：全球约 40% 岗位暴露	OpenAI / IMF
宏观 task-based，反方，	Acemoglu：十年 TFP≤0.66%、GDP 约 +1%	Acemoglu 在 2024 年
行业咨询综合	PwC：2030 年 AI 贡献全球约 $15.7 万亿	PwC 在 2017 年
全自动化→增长爆发	部分经济学家：年增速大幅跃升，具体结果高度依赖情景假设，争议较大	Davidson 等

数据来源：麦肯锡、OpenAI、IMF、Acemoglu、PwC 等公开研究，东吴证券研究所

5.2. 差异的根源

差异的根源有二：一是测的对象不同——是“现有生成式 AI 的渐进影响”，还是“AGI 全自动化带来的跃变影响”；二是关键假设不同——自动化的完整度、实体世界与监管的瓶颈、被替代劳动力的再配置速度。乐观估算（如麦肯锡、PwC）多假设自动化广而快、瓶颈小；审慎估算（如 Acemoglu）只计入“容易自动化任务”的成本节约，并认为剩余任务更难、可能高估。

6. 风险治理：审慎派与加速派并存

围绕 AI / AGI 的风险，业界存在两类对立鲜明的立场。本节按代表人物、核心观点与已采取措施分别梳理。

6.1. 审慎 / 风险派

Hinton 认为 AI 可能带来失控乃至生存性风险；Bengio 指出前沿模型已显现欺骗、作弊、黑客等危险能力；Stuart Russell 主张应转向“可证明有益（provably beneficial）”的 AI。已采取的措施：Hinton 于 2023 年离开谷歌以自由发声、呼吁立即监管与国际合作；Bengio 主持《国际 AI 安全报告》（96 位专家，2025/1），并创立非营利 LawZero（约 3000 万美元，safe-by-design）；Russell 创立伯克利 CHAI 与 IASEAI；Anthropic 推“负责任扩展政策”（RSP / ASL 分级），Ilya 的 SSI 以“安全超级智能”为唯一目标。政府与多边层面有欧盟《AI 法案》、英美 AI 安全研究所、Bletchley / Seoul 宣言。

6.2. 淡化 / 加速派

LeCun 认为生存性风险被夸大、当前 AI 并无自主意志；吴恩达认为 AI 恐惧被夸大（“担心火星人口过剩”），过度监管会扼杀创新；Marc Andreessen 在《技术乐观主义宣言》中视“AI 厄运论”为有害叙事。已采取的措施：LeCun 力推开源（Llama 系列）、反对过早过严监管；吴恩达倡导开源、反对牌照式监管；a16z 以资本与游说推动“Little Tech”议程、反对限制性监管。

中间 / 务实：微软 Suleyman 主张“人本超级智能、可控不自主”；多数政府走“分

级—透明—评测”的折中路线。

表8：AI 风险两派对照

阵营	代表	核心观点	已采取的措施
审慎/风险	Hinton	AI 或失控、生存性风险	离开谷歌发声、呼吁立即监管
审慎/风险	Bengio	前沿模型已现危险能力	主持《国际 AI 安全报告》、创 LawZero
审慎/风险	Stuart Russell	应转向“可证明有益”AI	创立 CHAI、IASEAI
审慎/风险	Anthropic / SSI	安全须优先于能力	RSP/ASL 分级；SSI 以安全为唯一目标
淡化/加速	LeCun	生存性风险被夸大	力推开源、反对过早监管
淡化/加速	吴恩达	AI 恐惧被夸大	倡开源、反对牌照式监管
淡化/加速	Andreessen a16z	“AI 厄运论”是有害叙事	资本+游说推 Little Tech
中间/务实	微软 Suleyman / 多国政府	可控不自主 / 分级治理	人本超级智能；分级—透明—评测

数据来源：各人物与机构公开资料，东吴证券研究所

两类立场的分歧，在于对“风险概率与时点”的判断不同，以及在“监管 vs 开源 / 创新”之间的权衡取向不同。

7. 风险提示：

技术迭代不及预期的风险：数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

大模型厂商 ARR 增速放缓的风险：若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

云服务商资本开支不及预期的风险：若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

智能驾驶及机器人商业化落地不及预期的风险：若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。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，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，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。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn