东吴证券 SOOCHOW SECURITIES

汽车行业点评报告

AI 系列深度 11 期：大模型如何进行模仿与强化学习？

2026 年 08 月 03 日

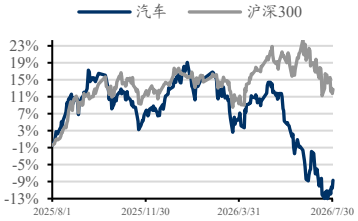
增持（维持）

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

投资要点

- **智能体学习行动主要有两条路径：**模仿学习，即从专家示范中学习状态到动作的映射；强化学习，即在环境交互中依据奖励信号优化策略。二者并非对立，模仿学习通常更快、更稳，强化学习在可构造奖励、仿真环境或可验证任务中更可能突破示范上限，工程实践往往采用先模仿形成初始策略、再强化优化的组合范式。
- **强化学习解决的是没有逐步标准答案、只有事后奖励反馈的问题，核心难点包括功劳分配、探索与利用、奖励稀疏和样本效率。**其思想源于行为主义心理学和贝尔曼 MDP，经 Sutton 与 Barto 体系化，并在 DQN、AlphaGo、RLHF、RLVR 和推理模型中持续演进。近年来，强化学习从补充性优化工具转向后训练、偏好对齐和可验证推理的重要方法。
- **模仿学习的本质，是把状态到动作的映射转化为监督学习。**行为克隆优点是训练快、稳定性较高，自动驾驶早期 ALVINN、机器人示教和 VLA 预训练都体现这一思路；局限在于分布漂移和示范者上限，模型一旦进入训练数据未覆盖状态，错误会逐步累积。逆强化学习、DAgger 和 GAIL 等方法，均围绕缓解这一问题展开。
- **在大模型训练中，强化学习主要用于后训练而非预训练。**预训练通过自监督学习通用表征，监督微调本质上是对高质量人工示范的模仿，RLHF 和 RLVR 再通过人类偏好或可验证奖励进行精修。AlphaGo 先学棋谱再自我对弈，ChatGPT 先 SFT 再 RLHF，机器人 VLA 先示范预训练再仿真或真实环境强化，均体现该组合范式。
- **落到智能驾驶和具身智能，纯模仿路线面临更强约束。**物理世界试错代价高，示范数据多来自平均人类表现，长尾和危险场景天然稀缺；因此趋势是用模仿学习获取初始策略，再依靠仿真、世界模型和强化学习补足长尾。杨立昆“蛋糕论”强调自监督是理解世界的主体，强化学习样本效率较低；但在可验证推理、偏好对齐和策略优化环节，强化学习的重要性正在上升。
- **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 10 期：预训练与后训练如何塑造大模型？》

2026-08-01

《AI 系列深度 09 期：从 LLM 到 VLM 再到 VLA 意味着什么？》

2026-07-31

内容目录

图表目录

1. 智能体学习行动：模仿与强化两条路径

智能体完成感知和表征后，需要基于当前状态选择行动。车辆规划驾驶动作、机械臂完成抓取、棋类程序选择落子，均属于决策与行动问题。大模型后训练中的强化学习同样建立在预训练形成的表征基础之上。因此，模仿学习与强化学习是从“理解世界”走向“采取行动”的关键环节。

同一行动学习问题对应两类基本路径。对于自动驾驶、棋类程序和机器人操作，模型需要学习从当前状态到动作选择的策略。模仿学习依赖专家示范，学习专家在不同状态下的动作；强化学习依赖环境反馈，智能体通过试错和奖励优化策略。

二者并不互斥，工程系统通常将其衔接使用。模仿学习能够快速获得初始策略，降低早期探索成本；强化学习则可在仿真、可验证任务或可构造奖励环境中进一步优化策略，尝试突破示范数据上限。

2. 强化学习：基于奖励反馈优化策略

2.1. 解决的问题：奖励、MDP 与三类难点

强化学习面对的是没有逐步标准答案、只有环境奖励反馈的决策问题。智能体（agent）在当前状态下选择动作，环境返回下一个状态和奖励（reward），智能体目标是在长期累计奖励最大化的方向上更新策略。这类问题通常用马尔可夫决策过程（MDP）描述，核心变量包括状态、动作、奖励和状态转移。

强化学习的三类难点较为典型：1）功劳分配，即任务结束后才看到结果时，如何判断哪些早期动作贡献更大；2）探索与利用，即继续采用已知有效动作，还是尝试可能更优的新动作；3）奖励稀疏，即多数时间步缺乏明确反馈，训练信号不足。

2.2. 方法源流：从心理学到现代算法体系

强化学习的思想可追溯至行为主义心理学。桑代克 1898 年的效果律指出，带来良好结果的行为更容易被强化；贝尔曼在 20 世纪 50 年代通过动态规划和马尔可夫决策过程奠定数学框架。Richard Sutton 与 Andrew Barto 进一步将强化学习体系化，提出时序差分（TD）学习等核心方法，并因相关奠基性贡献获得 2024 年图灵奖。此后，Q-learning、策略梯度、PPO 等算法相继形成，DQN 和 AlphaGo 则推动强化学习进入更广泛的产业视野。

2.3. 重要性提升：从补充工具到后训练关键方法

强化学习的重要性在近年明显提升。（1）AlphaGo 以自我对弈强化学习展示了机器在规则明确任务中突破人类示范上限的可能；（2）RLHF 使大模型能够根据人类偏好进

行对齐；（3）OpenAI o1、DeepSeek R1 等推理模型进一步使用可验证奖励强化学习提升数学和编程等任务表现。由此看，强化学习已成为后训练、偏好对齐和可验证推理的重要工具。但在数字 AI 中，它仍主要位于后训练和推理增强环节，并不替代预训练成为主引擎。

2.4. 大模型应用：主要位于后训练阶段

强化学习在大模型训练中主要用于后训练，而非预训练。预训练通常依靠自监督学习，通过下一个 token 预测等任务形成通用表征；监督微调（SFT）使用人工示范答案，本质上属于模仿学习；RLHF 进一步根据人类偏好优化模型输出，RLVR 则利用答案对错、测试是否通过等可验证奖励优化推理能力。整体看，预训练负责形成通用底座，强化学习主要承担对齐、纠偏和推理强化。

2.5. 样本效率争议：蛋糕论与客观边界

杨立昆的“蛋糕论”强调，自监督学习是学习世界表征的主体，监督学习和强化学习分别只是较小部分。其核心理由在于，强化学习单次交互得到的信息量有限，样本效率较低，难以单独从零建立对世界的理解。我们认为，该观点适用于表征学习阶段：从零获得世界模型和通用表征，主力仍是自监督学习；但在已经具备预训练底座后，强化学习在偏好对齐、可验证推理和策略优化中的作用正在上升。

3. 模仿学习：以专家示范形成策略

行为克隆是模仿学习的基础形式。模型将专家在不同状态下的观察和动作记录作为监督学习样本，直接学习从图像、传感器输入或状态到动作的映射。1989 年卡内基梅隆大学 ALVINN 使用三层神经网络从摄像头图像输出方向盘转角，被视为端到端神经网络自动驾驶的早期形态。

行为克隆的核心约束是分布漂移。模型训练时主要看到专家轨迹，一旦推理阶段因小误差进入训练数据未覆盖状态，后续动作可能继续偏离，误差逐步累积。Ross 与 Bagnell 在 2010 年对该问题进行了理论刻画。

缓解分布漂移主要有两类路线：1）逆强化学习（Ng 与 Russell，2000）不直接学习动作，而是从示范中反推专家优化的奖励函数，再据此求解策略；2）交互或对抗式模仿方法，例如 DAgger 让专家在模型偏离处补充示范，GAIL 则通过生成对抗学习使模型行为分布接近专家行为分布。

4. 两者分工：学习信号、风险与能力边界

模仿学习和强化学习的差异，首先来自学习信号不同，并进一步影响数据要求、训

练风险和能力上限。

表1：模仿学习与强化学习的多维对比

维度	模仿学习 Imitation Learning	强化学习 Reinforcement Learning
学习信号	专家示范，具备近似标准答案	环境奖励，事后反馈好坏，无逐步标准答案
数据	专家轨迹，可离线	在环境交互和试错中产生
主要软肋	走偏后误差会累积放大	奖励难设计、训练不稳、试错有代价
能力上限	通常难以超过示范者	在可构造奖励/仿真/可验证任务中有机会突破示范上限
速度与稳定	训练较快、稳定性较高	训练较慢、波动较大、算力消耗较高

数据来源：东吴证券研究所整理

整体看，模仿学习训练快、稳定性较高，但通常受限于示范者上限；强化学习在可构造奖励、仿真或可验证任务中具备突破示范上限的潜力，但训练较慢、稳定性较弱且试错成本更高。因此，二者在工程系统中常被组合使用。

5. 工程范式：先模仿形成初始策略，再强化优化

工程实践通常采用先模仿、再强化的两段式路径。（1）模仿学习用于快速获得可用初始策略，降低探索成本；（2）强化学习在可构造奖励或仿真环境中进一步优化策略，以突破示范数据上限或改善特定目标。

典型系统均体现该组合思路。AlphaGo 先通过监督学习吸收约三千万步人类棋谱，再通过自我对弈强化学习提升水平；ChatGPT 类大模型先使用人工示范回答进行监督微调，再通过 RLHF 对齐人类偏好；机器人 VLA 模型，如 Google RT-2、π0，主要依赖人类示范进行模仿式预训练，并在后续阶段引入仿真或真实环境优化。

逆强化学习提供了两类方法之间的桥梁。其核心是从专家示范中反推奖励函数，再交由强化学习优化策略，使模仿结果成为强化优化的起点。这种先形成初始能力、再进行目标优化的路径，与大模型预训练到后训练的分阶段范式具有相似性。

6. 落地到智能驾驶与具身智能：长尾决定强化需求

自动驾驶目前仍以模仿学习为重要基础。主流端到端方案通常从大规模人类驾驶数据中学习状态到动作的映射，使模型直接从画面和传感器输入学习驾驶策略。特斯拉 FSD v12 转向端到端神经网络并从人类驾驶视频学习；Wayve 系统结合模仿与强化学习。

真实物理世界限制了直接强化试错。自动驾驶和机器人任务中，一次错误可能带来车辆损坏或安全风险，真实数据采集成本也较高。因此，产业更多依赖仿真训练并迁移

至现实环境，但仍需面对仿真和现实之间的分布差异。

长尾场景是关键难题。纯行为克隆难以覆盖突然横穿行人等罕见危险场景，示范数据中的长尾样本天然稀缺，单纯增加数据规模也未必持续改善。因此，趋势是先用模仿学习建立基础策略，再用仿真、强化学习与世界模型补足长尾。

产业侧正在同时推进模仿学习与强化学习工具链。英伟达 Isaac 平台支持模仿与强化训练；大模型后训练方面，DeepSeek 提出 GRPO 并训练推理模型 R1，月之暗面、阿里、字节、腾讯、智谱、百度、MiniMax 等也发布或披露相关方案。整体看，路线仍在快速演化。

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。

东吴证券 SOOCHOW SECURITIES

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