东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 10 期：预训练与后训练如何塑造大模型？

2026 年 08 月 01 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

- **预训练与后训练是理解大模型能力形成的一组关键框架。**预训练利用海量低成本、弱标注或无标注数据学习通用底座，决定能力来源和上限；后训练利用少量高质量指令、偏好、反馈或可验证奖励数据进行适配，决定能力的调用方式、指令遵循、偏好对齐及特定任务中的推理表现。
- **当前模型能力与训练的关系主要聚焦三个问题：**能力主要来自预训练还是后训练，后训练究竟是塑形还是可带来新推理能力；两阶段是否正在扩展为预训练、中训练、后训练三阶段；规模化重心是否从预训练转向后训练与推理时计算。
- **数字 AI 与物理 AI 在训练框架上同构、在数据和反馈上分立。**数字 AI 预训练主要依赖文本和图文语料，后训练依赖人类偏好、可验证奖励和工具反馈；物理 AI 复用视觉语言骨干，但还需要驾驶视频、机器人轨迹、真机数据、仿真环境和安全奖励。公共文本可能接近供给上限，而物理 AI 的核心瓶颈是真实交互数据稀缺。
- **产业视角下，预训练仍属于资本开支密集的规模竞赛，后训练与推理效率则更多体现为以较小增量投入释放模型能力的效率竞赛。**随着推理模型、推理时计算和 Agent 应用发展，部分数字 AI 场景的算力重心正由训练侧向推理侧迁移；中国厂商也更多围绕后训练、推理效率、MoE、长上下文和工具调用开展工程优化。总体看，预训练决定能力底座，后训练成为重要增量来源，二者互补而非替代。
- **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 09 期：从 LLM 到 VLM 再到 VLA 意味着什么？》

2026-07-31

《AI 系列深度 08 期：GAN 与 Diffusion 两类生成范式有何差异？》

2026-07-31

请务必阅读正文之后的免责声明部分

内容目录

请务必阅读正文之后的免责声明部分

图表目录

请务必阅读正文之后的免责声明部分

1. 预训练与后训练：底座形成与使用适配

预训练与后训练共同构成大模型能力形成与产品化适配的核心链条。表征、视觉骨干、序列建模和生成模型提供了不同类型的数据理解与生成能力，但这些能力需要通过训练流程沉淀为可复用的模型参数。预训练在海量数据上以自监督方式学习通用表征，形成能力底座；后训练则面向人类意图、推理任务和使用边界进行适配，使模型从具备潜在能力转向可稳定调用。

核心定义：预训练（Pre-training）是指模型在海量、低成本、无需人工标注的数据上进行自监督学习，形成具备广泛知识与泛化能力的通用底座，决定模型的知识覆盖和能力上限；后训练（Post-training）是指在该底座之上，使用规模更小但质量更高、带有人类意图或反馈信号的数据进行适配，决定模型输出方式、指令遵循、偏好对齐和任务可用性。

从功能分工看，预训练侧重于知识、语法、模式和多模态关联的广泛吸收，使模型具备通用表达与生成基础；后训练侧重于将既有能力约束到可交互、可控和可评估的产品形态，使模型能够按指令回应、遵守安全边界并输出符合任务要求的结果。我们认为，这一“通用底座形成—使用场景适配”的两阶段范式，是理解数字 AI 与物理 AI 训练流程的主线框架。

1.1. 术语溯源：两个概念的提出与演变

预训练源于 2006 年的深度学习突破，本义是“逐层无监督初始化”：该词的技术起点可追溯至 Hinton、Osindero 与 Teh，以及 Bengio 等。当时深层网络难以直接训练，研究者先以无监督方式逐层初始化参数，再进行有监督微调，pre-training 指向正式训练前的参数准备。2015—2018 年，迁移学习路线赋予其今日含义：Dai 与 Le，Howard 与 Ruder，Radford 等，以及 Devlin 等，共同确立了“大规模数据自监督学习通用模型、再进行任务适配”的范式。

后训练在 ChatGPT 之后成为更通用的阶段性口径：在 GPT、BERT 时期，预训练之后的环节通常称为微调。InstructGPT 确立预训练、监督微调（SFT）、基于人类反馈的强化学习（RLHF）三阶段流程后，业界逐步将预训练之后、面向使用适配的训练环节统称为 post-training。此后该词被 OpenAI、Meta Llama 系列技术报告等沿用，外延覆盖 SFT、RLHF、RLAIF、DPO 直至 RLVR。后训练并非单一算法，而是一组面向可用性、可控性和推理增强的工序集合。

请务必阅读正文之后的免责声明部分

1.2. 本质区别：通用底座与面向使用的适配

表1：预训练与后训练的四维分工

维度	预训练 Pre-training	后训练 Post-training
数据	海量、低成本、无标注，全网文本、图像、视频	少量、高成本、高质量，人工示范、偏好排序、可验证答案
学习目标	自监督，以“预测下一个词”、掩码还原为主	模仿人类示范、对齐人类偏好、优化可验证奖励
信号来源	数据自身，无需人工评分	人类，标注、偏好，或环境，奖励、验证器
作用	形成通用底座，决定知识广度与能力上限	形成可用产品，决定指令遵循、格式、安全与能力的调用方式
规模特征	算力需求极大、周期长，主导训练成本	算力需求小、迭代快，对效果的边际撬动显著

数据来源：东吴证券研究所整理

由此可见，预训练决定模型内部能力的形成边界，后训练决定这些能力以何种方式被调用、约束和呈现。

1.3. 当前主要分歧：三条主线先行概览

当前分歧主要集中在三条主线：其一，能力主要来自预训练还是后训练，即表层对齐假说与强化学习可带来新能力两类观点的交锋；其二，训练流程应保持两阶段划分，还是扩展为预训练、中训练、后训练三阶段；其三，规模化投入的重心是否正从预训练转向后训练与推理时计算。现阶段，后训练与推理时计算的重要性提升在语言、数学、代码等可验证任务中体现更充分。

1.4. 数字 AI 与物理 AI：框架同构、口径分立

框架同构：物理 AI，英伟达等称之为 Physical AI，其代表性模型已明确沿用大语言模型训练配方。机器人基座模型 π0 被官方描述为预训练后接后训练、对标大语言模型范式；自动驾驶 VLA 模型英伟达 Alpamayo 则采用驾驶数据模态注入、因果链数据监督微调、强化学习后训练的多阶段流程。

但数字 AI 与物理 AI 在数据、反馈信号和核心瓶颈上存在明显差异：1）预训练数据不同，数字 AI 以文本中的预测下一个词为主，物理 AI 更多复用视觉-语言骨干，并在跨本体动作与驾驶视频上继续训练；2）后训练反馈信号不同，数字 AI 侧重人类偏好，物理 AI 侧重物理世界奖励、动作模仿与安全约束；3）核心瓶颈不同，数字 AI 面对公共数据接近上限，物理 AI 面对真机交互数据稀缺。因此，两者可共享分析框架，但具体训练路径仍需结合不同场景分别判断。

2. 预训练：通用能力底座如何形成

2.1. 自监督学习：以“预测下一个词”为核心目标

请务必阅读正文之后的免责声明部分

用数据自身提供监督信号：预训练的主力目标是自监督学习，即以数据内部结构作为监督信号，无需人工标注。典型路线包括预测下一个词，即根据上文预测后续词元；BERT 路线则采用掩码还原，即还原被遮挡的词元。我们认为，预测下一个词这一目标迫使模型在海量文本中学习语法、事实、常识和部分推理模式，进而成为通用能力的重要来源。

2.2. 扩展律 Scaling Law：规模与性能的定量关系

Kaplan 提出幂律、Chinchilla 校正配比：扩展律用公式刻画规模与性能之间的关系。Kaplan 等发现模型损失随参数量、数据量和算力增加呈幂律下降，意味着三者按一定比例扩大时，性能可沿相对可预测的曲线改善。Hoffmann 等通过更系统的实验加以校正：在给定算力下，参数与数据应大体等比例扩大，并提出约 20 个训练 token 对应 1 个参数的计算最优经验，意味着不少早期大模型存在参数偏大、数据偏少的问题。

2.3. 数据约束：高质量公共文本的供给上限

公共文本可能在 2028 年前后接近耗尽：Villalobos 等估计，可便捷获取的高质量公共文本将在约 2028 年基本用尽，早期估计为 2026 年；Sutskever 据此提出“我们只有一个互联网”，并将数据比作 AI 的化石燃料。应对路径包括：从追求更多数据转向追求更高质量数据、合成数据、向多模态扩展，以及将增量投入移向后训练与推理时计算。

物理 AI 预训练的数据来源有所不同：数字 AI 的预训练以文本或多模态场景中的预测下一个词为主；物理 AI 的预训练通常是复用已在网络数据上预训练的视觉-语言 VLM 骨干，再在跨本体动作数据或驾驶视频上继续训练，其本身已嵌套一次数字 AI 预训练。代表性数据如 Open X-Embodiment 汇集逾 100 万条轨迹、覆盖 22 种机器人本体。

3. 后训练：从基座可用到对齐与推理增强

3.1. 监督微调 SFT：习得指令遵循能力

用人工示范将续写能力转为应答能力：预训练得到的基座模型主要顺着上文续写，并不天然具备稳定的指令遵循能力。监督微调以人工编写的指令—回答示范为训练数据，使模型学习按指令作答、按预期格式输出，这是 InstructGPT 三阶段流程的第一步。

3.2. 基于人类反馈的强化学习 RLHF：对齐人类偏好

用偏好比较训练奖励模型、再优化策略模型：RLHF 的现代形态由 Christiano 等在 2017 年确立，流程是先让人类比较模型不同输出的质量，据此训练奖励模型，再用强化学习优化模型。其标志性结果是，仅约 900 比特的人类反馈，就让仿真机器人学会了连续动作后空翻。Stiennon 等在 2020 年将其用于文本摘要，InstructGPT 在 2022 年用其提升 GPT-3 的指令遵循、真实性和低毒性表现。

宪法 AI 与 RLAIF：以 AI 反馈替代部分人工标注：Anthropic 的宪法 AI 使用一组

请务必阅读正文之后的免责声明部分

人工制定的原则作为“宪法”，让模型自我评判并自动生成反馈；基于 AI 反馈的强化学习在保持效果的同时大幅减少人工标注，是后训练降本的重要路线。

3.3. 直接偏好优化 DPO 等简化路线

绕开奖励模型直接对齐偏好：Rafailov 等在 2023 年提出直接偏好优化，证明可绕过“训练奖励模型 + PPO 强化学习”的复杂流程，直接用偏好数据优化模型。相较 RLHF，DPO 训练更稳定、占用算力更少、也更不易出现“奖励欺骗”，已成为对齐的主流替代方案之一。

3.4. 可验证奖励的强化学习 RLVR：后训练的推理目标线

以可自动判定的奖励训练推理：当奖励可被自动验证时，例如数学题、代码可直接判定对错，无需人工逐一评分，即可大规模开展强化学习，称为 RLVR。DeepSeek-R1 采用 GRPO，甚至跳过监督微调、仅以强化学习便涌现出自我反思与验证等长链推理行为。

表2：后训练主要方法对照

方法	信号/数据	特点	代表
SFT 监督微调	人工示范，指令—回答	习得指令遵循与作答格式	InstructGPT 第一步，Ouyang 等，2022，
RLHF	人类偏好→奖励模型→强化学习	对齐偏好，更有用、更无害	Christiano 等 2017；InstructGPT 2022
DPO 等	人类偏好，直接优化	省去奖励模型，稳定、省算力	Rafailov 等，2023
RLAIF/宪法 AI	按“宪法”由 AI 自评反馈	降低人工标注成本	Bai 等 Anthropic，2022
RLVR	可验证奖励，数学/代码对错	激励推理、生成长思维链	DeepSeek-R1，2025

数据来源：东吴证券研究所整理

后训练的主要局限包括对齐税与奖励欺骗：对齐有时会牺牲部分原有能力，即对齐税；当模型学会迎合奖励信号而非完成真实任务时，会出现奖励欺骗。

4. 阶段过渡：中训练与推理时计算正在重塑边界

4.1. 中训练 mid-training：正在形成的第三阶段

中训练介于预训练与后训练之间，训练目标更为聚焦：中训练使用大规模无标注数据，但配以更聚焦的训练目标用于系统增强复杂能力并为后训练做准备；其中，数据退火是指逐步降低学习率以稳定收敛。phi-3.5 将其视为多语言与长上下文整合阶段，OLMo 2 强调学习率退火与数据配比，Yi-Lightning 视其为受控的数据分布过渡。

4.2. 推理时计算 test-time compute：推理阶段的算力投入

推理阶段引入额外计算：推理时计算指模型在回答时生成更多推理 token、进行更长步骤的中间推演，以推理侧算力换取更高准确率。Snell 等在 2024 年指出，最优地分

请务必阅读正文之后的免责声明部分

配推理时计算，在许多任务上比单纯扩大参数更有效。OpenAI o1 在 2024 年以更长思维链换取更强推理表现，DeepSeek-R1 也证明纯强化学习可使模型在推理阶段生成更长的中间过程。

4.3. 思维链 CoT：横跨预训练、后训练与推理的连接点

思维链是一种推理行为，而非独立训练阶段：思维链指模型在给出最终答案前，先输出中间推理步骤，再形成结论。思维链本身不是一个训练阶段，而是一种可在推理时呈现、可在后训练中被强化的行为模式；它横跨预训练、后训练与推理三处，是理解两阶段范式被重塑的典型案例。

能力基础来自预训练，可在推理阶段被触发：模型能够输出有条理的推理步骤，源于预训练语料中包含大量分步推理文本；Wei 等与 Kojima 等证明，只需在提示中给出分步示范或一句引导语，即可在推理阶段触发该能力，无需额外训练。

在后训练中被放大，并成为推理时计算的载体：o1 与 DeepSeek-R1 在 2025 年进一步用强化学习 RLVR 让模型无需提示便生成更长、更有效的思维链，把推理行为从提示诱发转向默认输出；部署阶段中，这种长思维链正是推理时计算的具体形式。我们认为，思维链由此成为连接后训练与推理时计算的关键机制。

5. 数字 AI 与物理 AI：同一框架，不同数据和反馈

5.1. 机器人：复用视觉-语言骨干预训练，依托真机与仿真后训练

从 RT-2 到 π0：RT-2 在视觉-语言模型 PaLM-E、PaLI-X 之上，联合微调机器人轨迹数据与网络图文数据，并把动作表示为文本 token。π0 以 PaliGemma 为骨干，预训练阶段使用逾 10000 小时、覆盖 7 类机器人 68 项任务的数据并叠加 Open X-Embodiment，再以高质量数据后训练，并采用流匹配方法生成连续动作轨迹，输出高频动作。机器人领域的核心分歧之一，是动作应表示为离散 token 还是连续数值。

5.2. 自动驾驶：从端到端到 VLA，后训练引入强化学习

UniAD、世界模型与 Alpamayo：UniAD 将感知、预测、规划统一进一个可微分的端到端框架。世界模型用自回归 Transformer 加扩散解码生成未来驾驶视频，常被置于训练、仿真、评估和策略改进基础设施一侧，而非简单归入预训练。英伟达 Alpamayo 则将大语言模型配方移植到车端：先进行大规模驾驶数据模态注入与预训练，再通过因果链数据监督微调习得推理，最后以强化学习后训练优化轨迹安全与推理-动作一致性。

5.3. 物理 AI 的核心瓶颈：真机数据的稀缺

数据规模而非术语是分水岭：即便 Open X-Embodiment 汇集了百万级轨迹，相对文本语料仍属小数据；高质量真机数据稀缺且昂贵，遥操作采集既慢又易出错，成为部署与扩充数据的实际瓶颈。由此，数据来源分化出三条路线：海量人类操作视频、真机

请务必阅读正文之后的免责声明部分

采集、仿真。

6. 产业与投资视角：成本与价值的迁移

6.1. 算力重心：从训练转向推理

推理正成为 AI 经济的重要成本中心：随着基座模型成熟、商用部署铺开，AI 的成本重心在多个数字 AI 场景中从训练转向推理。训练是一次性固定投入，推理则是被高频调用的持续边际成本。业界估计，2025 年主要厂商的推理算力占比已过半。黄仁勋提出“推理拐点已至”，并指出具备推理能力的模型既要生成更多 token、又要更快响应，叠加后对算力的需求可达约 100 倍。

6.2. 预训练的规模红利与后训练的效率红利

两阶段对应两种竞争逻辑：预训练是资本开支密集的规模竞赛，数据与算力门槛高，由少数基座厂商主导；后训练则是以较小增量投入提升模型可用性和推理能力的效率竞赛。DeepSeek-R1 采用 GRPO 并跳过监督微调，以远低于同类的成本逼近 o1 级推理，并可将能力蒸馏迁移到更小规模。我们认为，在数据接近上限、预训练边际收益趋缓的预期下，后训练与推理效率正成为新的增长极与差异化来源。

6.3. 中国厂商的布局：聚焦后训练与推理效率

以效率与性价比切入：我们梳理认为，中国头部厂商更多在后训练算法与推理效率上发力。DeepSeek 围绕 R1、GRPO 与能力蒸馏形成代表性样本；阿里 Qwen 系列、字节豆包、月之暗面 Kimi、MiniMax、智谱 GLM 等也密集迭代推理模型，其中 GLM-5 公开口径称在华为昇腾平台训练。整体看，国内厂商更多以后训练与推理优化换取性价比。

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。

请务必阅读正文之后的免责声明部分

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。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

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