后训练，RL，左脚踩右脚

人工智能的自我进化之路：从预训练到强化学习驱动的能力跃迁

下面是一些杂谈

一、 基础：预训练与知识的建立

大语言模型（LLM）开发的第一步，是进行大规模的预训练。这个过程的核心，是将模型暴露在极其庞大的文本数据集之中，这些数据通常来源于互联网、书籍等多种来源，覆盖了人类知识和语言使用的广泛范围。模型通过无监督学习的方式处理这些数据，最常见的任务是预测文本序列中的下一个词，或者填补文本中被掩盖的部分。通过完成这些看似简单的预测任务，模型得以在内部构建起对语言的复杂理解。

在预训练过程中，模型不仅仅是记忆文本片段，更重要的是学习到了语言的底层结构，包括语法规则、词语搭配、上下文依赖关系。同时，它也吸收了数据中蕴含的大量事实性知识，以及概念之间微妙的语义联系。可以说，预训练完成后的模型，内部形成了一个关于世界的庞大统计模型——它“知道”了构成语言的符号是如何组合的，以及这些符号组合在描述世界时呈现出的模式和规律。GPT系列模型的发展历程，特别是从 GPT-2 到 GPT-3 的跨越，清晰地展示了通过增加模型参数规模和预训练数据量，能够显著提升模型掌握这种基础知识和语言能力的深度与广度。因此，预训练的目标是为模型打下一个坚实的知识基础，使其理解“世界是什么样子的”，至少是文本数据所描绘的那个世界的样子。

然而，预训练本身并不直接赋予模型执行特定任务或进行符合人类期望交互的能力。模型虽然储存了海量信息，理解了语言模式，但它本质上是根据统计概率生成文本，缺乏明确的目标导向。它不知道在具体的对话情境下，什么样的回答是“好的”、“有用的”或“安全的”。它可能生成事实准确但毫无帮助的回答，或者在没有明确指令的情况下延续不相关的话题。简而言之，预训练让模型具备了“知其然”的潜力，但并未教会它“知其所以然”的应用智慧，即如何根据特定目标或用户意图来组织和运用其庞大的知识储备。这种内在能力的缺失，使得仅经过预训练的模型难以直接投入实际应用，也自然引出了模型开发的下一个关键阶段：后训练。

二、 对齐：引导模型理解并遵循人类意图

预训练赋予了模型广博的知识基础，但正如前述，这并不足以让模型在实际应用中表现良好。模型需要进一步学习如何运用这些知识来响应用户的具体需求，以及如何使其行为符合人类的价值观和偏好。这个过程通常被称为**后训练 (Post-training)** 或**对齐**

(Alignment)，其核心目标是弥合模型固有能力与用户期望之间的差距，教会模型“知道应该怎么做”。

后训练的第一步往往是**监督微调 (Supervised Fine-Tuning, SFT)**。在此阶段，模型学习模仿由人类专家精心编写或筛选的高质量“指令-理想回答”范例。SFT 的主要作用是初步塑造模型的行为模式，让它熟悉交互的基本格式，并掌握遵循明确指令的能力，为模型提供了一套“行为模板”。

随后，为了让模型具备更灵活、更可靠的对齐能力，并更好地处理依赖主观判断的情况，研究者们引入了**基于人类反馈的强化学习 (Reinforcement Learning from Human Feedback, RLHF)**。通过让人类对模型生成的多个回答进行排序或评分，训练一个**奖励模型 (Reward Model)** 来模拟人类偏好，再利用强化学习算法（如 **PPO**）根据奖励信号微调语言模型，鼓励其生成更符合人类期望的内容。

值得注意的是，深度学习领域一个引人瞩目的现象在于其**目标驱动的学习特性**：当你以恰当的方式要求模型学习一个复杂的任务时，它往往会为了达成这个目标而自发地学习掌握完成该任务所必需的各种潜在能力。SFT 和 RLHF 正是利用了这一点。当我们要求模型学习“如何像人一样有效对话”——一个本身就极其复杂的任务时，模型为了获得好的 SFT 效果（精确模仿高质量范例）或高的 RLHF 奖励（生成人类偏好的回答），就**不能仅仅停留在表面模仿或简单的模式匹配上**。

有效的对话天然地要求理解上下文的细微差别、准确把握用户意图、从自身知识库中检索并组织相关信息、进行一定程度的逻辑推理以确保回答的连贯性和合理性，并最终以清晰、有条理的语言表达出来。因此，模型在努力“回答得更好”的过程中，实际上被**“倒逼”着去发展和优化这些更深层次的、类似人类思维的工作方式**。这个看似主要目标是“行为对齐”的训练过程，其客观结果却是**极大地促进了模型通用“智力”的提升**，增强了它在更广泛问题上的解决能力。模型为了“做得像人”，不得不“想得更像人”。这在很大程度上解释了为何经过后训练的模型，相比其预训练的“基座”版本，在理解力、响应相关性、甚至初步的创造性任务上都表现出显著的飞跃。可以说，**对齐训练不仅规范了模型的言行，也无意中挖掘并强化了其认知潜力**，使其更加接近一个真正有用的、通用的智能助手。

这种伴随对齐而来的智能提升，为模型能够胜任更多样化的任务打下了坚实的基础。值得注意的是，这种提升主要体现在模型**处理相对短时、信息明确的交互和问答**的能力上。因为 SFT 和 RLHF 的训练数据和评估方式，往往聚焦于单个回合或有限轮次的对话质量，以及对给定信息的即时响应。模型因此精通于在形式上进行连贯的问答，或者在给定上下文中进行信息整合与知识检索。而这种训练模式对于培养模型进行**长期规划、处理复杂依赖关系、或在需要持续探索和适应的任务中保持目标导向**的能力，则显得力有不逮。这就导致了我们观察到的现象：模型可能在某些封闭的、知识密集型的任务（如特定领域的数学题）上表现出色，因为它本质上是在执行一种复杂的模式匹配和符号操作；但在那些需要理解深层逻辑、进行多步规划、或者处理新颖情境的任务（哪怕是看似简单的小学应用题或需要常识推理的谜题）上，却可能暴露

出明显的短板。这种“智力”上的不均衡，根源在于当前的对齐训练更多地强化了模型的“应答”能力，而非“解决问题”的深层策略能力。

因此，虽然对齐极大地提升了模型的通用性和交互性，但当研究者们追求让模型在更复杂、更需要深度思考和长远规划的任务上取得突破时，这种主要面向“对话优化”的训练所带来的智能提升，就遇到了它的边界。这就要求研究者去探索新的方法来进一步强化模型的核心认知能力，特别是那些超越了简单问答模式的、更为深刻的推理和规划能力。

三、 挑战与启示：推理能力的瓶颈与思维链（CoT）的曙光

正如前文所述，虽然通过 SFT 和 RLHF 进行的对齐训练极大地提升了模型的通用交互能力和泛化智能，但当研究者们将目光投向那些需要更深层次、更严谨逻辑的**复杂推理 (Complex Reasoning)** 任务时，便触及了现有方法的明显瓶颈。研究者们很快发现，尝试直接运用或扩展标准的 RLHF 框架来专门强化模型的推理能力，面临着一系列严峻的挑战。

首当其冲的是奖励信号的设计与应用难题。 对于一个需要多步骤才能完成的推理问题（例如复杂的数学应用题或逻辑谜题），如何设定一个有效指导模型学习的奖励机制本身就极其困难。最终答案的正确与否虽然是一个清晰的信号，但它过于稀疏，无法为漫长推理链条中的每一个决策点提供及时的反馈。奖励模型很难精确地判断某个中间步骤是对是错，或者哪种思考路径更有潜力。这种**奖励信号的模糊性与稀疏性**导致了棘手的**信用分配 (Credit Assignment)** 问题：模型难以知道究竟是哪一步的“神来之笔”或“致命失误”导致了最终的成功或失败，从而使得基于奖励的学习效率低下。

其次，奖励模型本身的可靠性也备受考验。 即便奖励模型能够对最终输出给出评分，它也可能存在自身的认知偏差，或者更容易被模型利用策略“钻空子”（即 Reward Hacking）。模型可能学会生成形式上看似合理、符合奖励模型偏好但逻辑上存在谬误的推理过程，而非真正提升内在的逻辑推理能力。

此外，推理任务的广阔探索空间也给强化学习带来了困难。 复杂问题的解空间往往巨大且结构复杂。让模型在缺乏明确引导的情况下，单纯依靠“试错”来探索出一条有效的长推理路径，不仅需要巨大的样本量和计算资源，而且效率低下，也容易陷入局部最优。同时，模型很可能只是过拟合了训练数据中特定类型的推理模式，难以泛化到新的、结构相似但细节不同的问题上。

正当研究界在积极应对这些利用强化学习提升推理能力的挑战，寻求更有效训练范式之际，一项起初看似简单、主要发生在**推理阶段 (Inference-Time)** 的发现，却带来了意想不到的曙光，并深刻地改变了人们对大模型潜力的认知。这就是**思维链 (Chain-of-Thought, CoT) 提示法**的出现。大约在 2022 年，研究者们注意到，仅仅通过在输入提示中加入简单的引导语（如“让我们一步一步地思考”）或包含少量带有推理步骤的

示例，就能**显著提升大语言模型在需要多步推理任务上的表现，而无需对模型进行任何额外的训练或参数调整**。

CoT 的核心机制是诱导模型在输出最终答案前，先**显式地生成一系列中间思考步骤**，如同人类解决问题时的草稿或自述分析。这种方式似乎能够帮助模型更好地组织和运用其内部已有的知识与计算能力，将复杂问题分解为更易于处理的子步骤，从而更准确地完成推理。

CoT 的成功带来了极其重要的启示：

- 它力证了**大型语言模型内部蕴藏着远超人们先前预期的推理潜力**，只是这种潜力需要被恰当的方式“解锁”或引导出来。
- 它凸显了**“思考过程”本身对于解决复杂问题的重要性**。输出一个逻辑清晰、步骤合理的推理链，不仅是通往正确答案的有效途径，也使得模型的决策过程更加透明、可被理解和调试。
- 最关键的是，它为**模型训练设定了新的努力方向**。既然模型 *能够* 在引导下生成有效的推理过程，那么核心问题就转变为：我们能否设计出**更有效的训练方法**，让模型**内在地、稳定地掌握这种生成高质量推理链的能力**？能否让模型摆脱对特定提示技巧的依赖，真正将结构化思考的能力固化为自身的核心竞争力？

因此，CoT 的出现，虽然没有直接提供解决 RLHF 困境的训练方案，但它极大地激发了研究热情，并清晰地指明了未来探索的方向——即训练的重点需要**从仅仅关注最终输出的质量，转向更加关注和优化模型生成完整、可靠思考过程的能力**。这为后续旨在直接提升模型核心推理智能的新型强化学习算法（如 GRPO）的诞生，奠定了重要的认知基础和研究动机。

四、 突破：从引导思索到内化智能

思维链（CoT）的发现如同一把钥匙，打开了我们对大模型推理潜力的新认知，关键在于显式化其“思考过程”。但这把钥匙最初只是在模型“门外”使用（推理时提示），真正的挑战是如何将这种能力“装进门内”，让模型**自主地、内在地掌握高质量的链式思考能力**。这需要训练方法的根本性突破，以克服传统 RLHF 在优化复杂推理时遇到的种种障碍。强化学习领域为此展开了关键的探索，并取得了决定性的进展。

实现这一目标，并非单一技术能够完成，而是往往需要一个精心设计的、循序渐进的训练策略。一个被证明有效的思路是，首先通过**监督微调 (SFT)**，为模型搭建起进行结构化思考的“脚手架”。在这个阶段，重点可能并不仅仅是让模型模仿正确的答案，更是**教会它采纳一种能够清晰展示其思考过程的输出格式**——例如，学习使用特定的标记（如 <think>...</think>，以及使用各种转折/汇总/推理词汇，如“也许”、“但是”、“因此”）来界定其内部的推理步骤。这相当于给模型一个明确的信号和规范：“当你需

要解决复杂问题时，请在这里、以这种方式展示你的思考”。这利用了深度学习模型强大的模式学习能力，让它先学会“在哪里”以及“如何”表达思考。

一旦模型掌握了这种结构化的表达方式，即搭建好了“思考的框架”，下一步的关键就是**提升填充在这个框架内的“思考内容”的质量、逻辑性和有效性**。这正是强化学习（RL）发挥其核心作用的阶段。仅仅依靠 SFT 的模仿，模型生成的思考过程可能仍然流于表面或充满谬误。而引入 RL，特别是像**组相对策略优化 (Group Relative Policy Optimization, GRPO)** 这样的先进算法，则能够更有效地引导模型向更高水平的推理能力迈进。

GRPO 负责了 R1 模型训练的强化学习部分，在此过程中显示出了独特价值，它通过一种不同于传统 PPO 的机制来提供学习信号。具体而言，当模型接收到一个需要推理的输入提示时，GRPO 会**首先驱动模型基于其当前策略生成一整组（例如 N 个）不同的完整输出序列**。每一个序列都包含了由 SFT 阶段教会的、用于表达思考过程的结构（如 <think>...</think> 标签）以及最终的答案。

接下来，一个**奖励模型 (Reward Model)** 会介入，**对这一组中的每一个完整输出序列进行独立的打分评估**。这个奖励模型被特别设计或训练用来判断整个思考过程的质量——包括逻辑的严谨性、步骤的合理性、知识的准确运用，以及最终答案的正确性等多个维度。它着眼于整体解决方案的优劣，**而非试图给中间步骤定价**。

然后，GRPO 执行其核心操作：**计算这一组 N 个输出得分的平均值，并将这个平均值作为当前批次的动态判断基准 (Baseline)**。有了这个基准，对组内任何一个具体的输出序列而言，其学习信号（即**优势 Advantage**）就通过**比较自身得分与这个组平均得分的差值**来获得（即 Advantage = Reward_i - Average_Reward_Group）。

这个“相对优势”信号非常关键：它直接告诉模型，其这一次生成的特定思考路径和答案，相比于它当前能力下为同一个问题生成的其他可能方案，是更好还是更差。**GRPO 随后利用这个相对优势信号来更新模型的策略**：对于那些得分高于组平均的、被认为是“相对更好”的输出序列，模型会调整参数以增加未来生成类似高质量推理和答案的概率；反之，则降低生成低于平均水平的输出的概率。通过这种方式，GRPO 高效地将奖励模型对整体解决方案质量的评估，转化为了对模型策略的具体优化指导，实质上是在激励和引导模型学习如何在之前由 SFT 建立的框架内“思考得更好”。

值得强调的是，模型推理能力的提升，特别是生成更长、更复杂思考链的现象，并非 GRPO 算法直接“规定”的结果。更准确地说，这是模型在追求更高奖励（即解决更复杂问题、展现更优逻辑）的过程中，**自发学习和“涌现”出的能力**。当模型发现更详尽、更深入的思考步骤有助于获得更高评价时，它自然会倾向于这样做。这再次体现了深度学习的内在机制：设定好正确的学习目标和反馈机制，模型就能在训练中展现出惊人的适应性和能力增长。

因此，GRPO 等先进 RL 方法的成功，并非孤立的算法创新。它们是**在一个结合了结构化引导（SFT）和质量优化（RL）的整体训练策略中，扮演了关键的“引擎”角色**，最

终实现了将 CoT 式推理能力从外部技巧转化为模型**内在核心技能**的重大突破。以 DeepSeek-R1 等模型的表现为例，它们证明了这种训练范式的有效性，标志着我们真正掌握了利用强化学习来**直接、深入地培育和提升模型核心认知能力**的手段。**这一成就，不仅为攻克 AI 在复杂推理上的瓶颈铺平了道路，更关键的是，它有力地证明了强化学习驱动智能进化的巨大潜力，从而将我们带入了一个全新的、以 RL 为核心引擎的 AI 发展阶段（这里主要指大语言模型）**。

五、 步入强化学习驱动的新时代

（以下两节都是在大语言模型的背景下进行讨论，对于之前的 RL 技术，请参照 AlphaGo 的发展历程）

随着像 GRPO 这样的先进强化学习技术展现出其在提升模型核心认知能力（尤其是推理能力）上的强大潜力，大语言模型的发展，正无可争议地迈入一个以**强化学习（RL）为核心驱动力的新时代**。这个时代的到来，其重要性远不止于算法本身的进步，它从根本上改变了我们看待和发展人工智能的方式。

在此之前，模型能力的提升主要依赖于更大规模的数据和模型（预训练），以及基于人类示范和偏好的外部引导（SFT 和早期 RLHF）。这些方法卓有成效，但本质上仍是将模型视为一个被动学习的对象。然而，强化学习，特别是当它被成功应用于优化如逻辑推理、数学解题、代码生成这类复杂的内部“思维过程”时，情况发生了质变。RL 提供了一种**目标导向的、持续迭代的优化框架**。模型不再仅仅是吸收静态知识或模仿固定行为，而是能够在与环境（或其奖励模型代理）的动态交互中，通过“试错”和反馈，主动探索和学习如何达成复杂的目标。

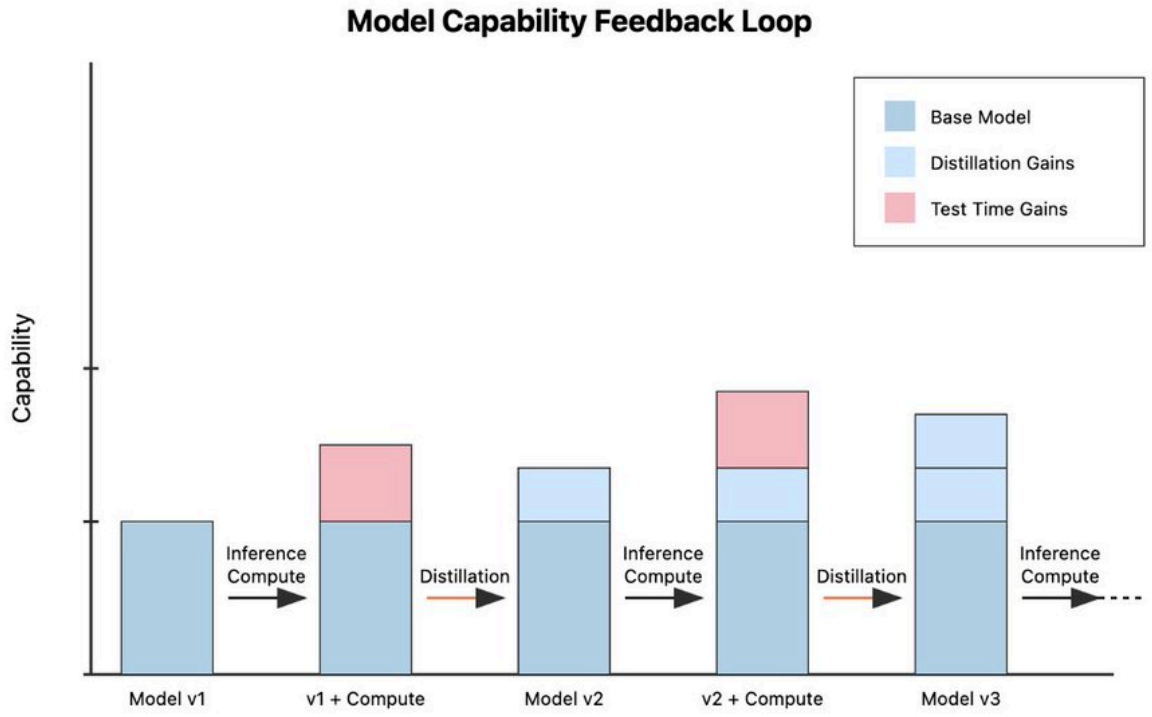
这个转变的关键在于，**模型本身的核心智能——它的推理能力、规划能力、甚至未来可能发展的自我评估与反思能力——成为了可以直接通过 RL 进行塑造和强化的对象**。当一个模型的推理引擎可以通过 RL 变得更强大、更可靠时，它就不再仅仅是一个被动的知识容器或指令执行者，而更像是一个可以被持续“打磨”和“升级”的智能核心。这种可能性是革命性的，因为它意味着模型能力的提升不再完全受限于外部数据的质量和规模，或者人类指导的效率和成本。

强化学习时代的开启，为人工智能的未来发展描绘了一幅激动人心的蓝图。它不仅有望解决一些长期困扰我们的难题，比如如何让 AI 在开放、复杂的环境中做出真正智能的决策，更重要的是，它为一种更深层次的进化——AI 的自我完善——打开了大门。当模型的核心能力可以通过 RL 得到增强，这些增强后的能力本身，又可能反过来被用作进一步提升模型性能的工具或资源。这种内在的、潜力巨大的正反馈循环，正是我们接下来要探讨的“左脚踩右脚”式自我进化机制得以生根发芽的土壤。可以说，**强化学习时代的到来，不仅是技术路线图上的一步，更是通往更强大、甚至可能实现自主进化的人工智能的关键里程碑**。

六、 自我进化机制：“左脚踩右脚”的实践路径

步入强化学习驱动的新时代后，人工智能的发展展现出一种前所未有的可能性：模型不再仅仅依赖外部输入，而是能够利用其自身不断增强的能力来驱动进一步的成长。这种被形象地称为“**左脚踩右脚**”的自我进化机制，本质上是一种自举（Bootstrapping）或自我完善（Self-Improvement）的过程。它标志着 AI 从主要由人类设计和数据喂养，开始转向探索由自身能力驱动的、潜力巨大的内生性增长。根据当前的研究进展和工程实践，这种自我进化可能主要通过以下两条关键路径实现：

路径一：以内在思考指导未来学习——测试时计算增强与蒸馏



这条路径的核心思想是，充分利用模型在**特定条件下能够达到的更高智能水平**，并将这种“巅峰表现”转化为可供下一代模型学习的宝贵经验。这里的“特定条件”，主要是指在模型进行**推理或响应（即“测试时”）**，通过投入额外的计算资源，引导其进行更深入、更结构化的思考。

我们已经知道，像思维链（CoT）这样的技术，可以在不改变模型参数的情况下，仅通过提示工程，就让模型在解决复杂问题时表现得更好。更进一步，研究者们开发了多种**测试时计算增强 (Test Time Compute, TTC)** 技术，例如多路径采样与自洽性检验（生成多个推理路径并择优）、自我反思与迭代细化（让模型批判和修改自己的初步答案）等。这些技术本质上都是让模型在回答问题前“多想一会儿”、“想得更周全”，从而显著提升输出的质量和准确性。这种通过额外计算换来的性能提升，可以看作是模型在特定时刻被激发出的、超越其基础能力的“**测试时增益 (Test Time Gains)**”。

然而，这种增益是临时的，依赖于推理时的额外计算开销。那么，能否将这种“灵光一闪”或者“深度思考”的能力，**永久地固化到模型的基础能力中呢？** 这就是**知识蒸馏 (Knowledge Distillation)** 发挥作用的地方。具体做法是：收集由现有模型（比如 Model v1）通过 TTC 技术生成的大量高质量输出——这些输出不仅包括最终答案，最好还能包含详细、准确的推理过程或思考步骤。然后，将这些高质量的“思考样本”作为“教师”的示范，用来**训练或微调下一代模型（Model v2）**。

在这个蒸馏过程中，Model v2 的学习目标不再是简单地模仿原始、可能质量不一的训练数据，而是去学习其“前辈”（v1 + Compute）在高计算强度下所展现出的**更高水平的思考模式和解决问题的能力**。通过这种方式，原本需要额外计算才能获得的“测试时增益”，就有可能被部分地**内化 (Internalize)** 到 Model v2 的参数之中，成为其新的、更高的基础能力（这部分提升可称为“**蒸馏增益 Distillation Gains**”）。

这个“**TTC 增强 -> 蒸馏内化**”的过程可以形成一个持续迭代的**模型能力反馈回路 (Model Capability Feedback Loop)**。每一代模型都在前代模型“深度思考”成果的指导下进行学习，从而实现基础能力的不断提升。这不仅加速了模型智能的进化，也间接提高了对原始数据价值的利用效率——模型不再是被动消化，而是主动地对知识进行“精加工”并传承下去。强化学习技术，如 GRPO，也可以在这个蒸馏/训练环节扮演重要角色，例如通过 RL 优化 Model v2，使其能更好地生成符合“教师”模型高质量推理过程的输出。

路径二：以自我评估驱动持续优化——AI 反馈强化学习 (RLAIF)

如果说第一条路径是利用模型增强后的“思考能力”来指导学习，那么第二条路径则是利用模型可能发展出的“**评估与判断能力**”来驱动自身的优化，这主要体现在**基于 AI 反馈的强化学习 (Reinforcement Learning from AI Feedback, RLAIF)** 上。

传统的 RLHF 严重依赖人类标注者提供偏好反馈，这不仅成本高昂、速度缓慢，而且人类判断本身也可能存在不一致性或偏见，这些都限制了 RLHF 的规模和应用范围。RLAIF 的核心思想是用一个（或多个）**训练有素的 AI 模型来代替人类评估者**，充当**奖励模型 (Reward Model)** 或“**AI 裁判**”。这个 AI 裁判的目标是学习模仿人类的评价标准，或者遵循预先设定的原则（如 Constitutional AI），来判断其他模型输出的质量、安全性、有用性等。

一旦我们拥有了一个足够可靠、高效的 AI 奖励模型，整个强化学习的流程就可以实现大规模的自动化。待优化的语言模型可以自由地生成大量的候选输出；AI 奖励模型则快速地对这些输出进行评估，并给出相应的奖励分数或偏好排序；这些由 AI 产生的反馈信号随后被用于驱动强化学习算法（如 PPO, GRPO, 或 DPO 等），持续优化语言模型本身的策略。

这个“**模型生成 -> AI 评估 -> 模型学习**”的闭环，正是 RLAIF 的精髓所在。它构成了一个强大的自我驱动循环：模型的能力提升可以帮助训练出更好的 AI 奖励模型（因为它

能提供更高质量的评估样本或自身就具备更强的判断力），而更好的 AI 奖励模型又能更有效地指导模型的进一步优化。这极大地**突破了人类反馈的瓶颈**，使得强化学习可以在更广阔的范围内、以更快的速度进行迭代，从而加速模型的对齐过程和特定能力的提升。例如，DeepSeek 提出的 **GRM (Generalized Reward Model)** 概念，不仅让 AI 能打分，还能生成评价原则和批评，为 RLAIF 提供了更丰富的反馈信息。

总而言之，这两条路径——无论是通过“TTC+蒸馏”实现能力的内化传递，还是通过“RLAIF”实现自我评估驱动的持续优化——都体现了“左脚踩右脚”式自我进化的核心理念。它们都依赖于模型**利用自身已经获得或发展出的高级能力（思考能力或评估能力），在强化学习等机制的驱动下，实现进一步的、可能加速的自我完善**。这标志着 AI 发展进入了一个潜力无限的新阶段，模型不再仅仅是工具，更可能成为自身进化的主动参与者。