强化学习的奠基人的惊人警告：为什么说LLM可能是一条死胡同?

2025年10月2日 1 分钟阅读

#AI #Richard Sutton #LLM #RL #Goal-Driven #Continual Learning

> *强化学习的奠基人惊人警告：为什么说LLM可能是一条死胡同?*

在最近的一次Dwarkesh Patel访谈中（2025-09），强化学习（RL）的奠基人之一理查德·萨顿（Richard Sutton）对大型语言模型（LLM）提出了尖锐的批评，认为它们是AI研究的"死胡同"。萨顿认为，智能的核心在于实现目标和从经验中学习的主动过程，而LLMs缺乏目标和地面实况（ground truth），仅是模仿人类行为，这与RL范式形成了鲜明对比。他阐述了RL中值函数、策略和世界模型的组成部分，并强调了持续学习和从经验中学习的必要性，认为这才是通往通用人工智能的可扩展方法。此外，萨顿还讨论了他著名的"痛苦的教训"（The Bitter Lesson）理论，以及他对AI继承和人类社会向设计智能过渡的宇宙视角的看法。

> *Richard Sutton 是强化学习的奠基人之一，发明了时间差分学习（TD learning）、策略梯度方法等核心技术，获 2024 年图灵奖（计算机领域的 "诺贝尔奖"），此次对话聚焦从 RL 视角剖析 AI 领域对 LLMs 的认知局限及 AI 发展的关键方向。*

引言：超越狂热，重审智能的本质

在人工智能领域，几乎所有的对话都被一个话题所主导：大语言模型（LLM）。从初创公司到科技巨头，整个行业似乎都沉浸在LLM带来的惊人进展和无限可能性之中。它们生成文本、编写代码、回

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #Richard Sutton #LLM #Goal-Driven #Continual Learn

答问题的能力达到了前所未有的水平，以至于很多人相信，通往通用人工智能（AGI）的康庄大道，就是建造越来越庞大的模型。

然而，在这股压倒性的浪潮中，一个响亮而有力的不同声音出现了。他就是理查德·萨顿（Richard Sutton），一位图灵奖得主，被誉为“强化学习的奠基人”之一。萨顿的观点并非温和的质疑，而是一种根本性的挑战。他认为，当前以LLM为中心的研究路径可能是一条“**死胡同**”。萨顿的批判并非零散的观点集合，而是源于一个统一且坚定的核心信念：智能是一个智能体（agent）通过与环境直接互动来学习，从而实现其目标的计算能力。

这篇博文的目的，就是深入剖析萨顿基于这一核心原则所提出的一系列颠覆性论点。这些论点共同构建了一个与主流叙事截然不同的AI世界观，它将引导我们超越模仿的表象，回归学习与智能的真正本质。

1. 核心论点：LLM是“模仿者”，而非真正的“学习者”

萨顿对智能的根本定义，直接引出了他对LLM范式的第一个核心批判：LLM本质上是“模仿者”，而非“学习者”。他指出，大语言模型从一个静态的、巨大的文本语料库中学习——也就是人类已经说过的话。其核心功能是基于这些数据，预测一个普通人在特定情境下接下来会说什么。

这与萨顿所定义的、源于经验的真正学习形成了鲜明对比。真正的学习是一个智能体在世界中执行一个动作，观察该动作带来的结果，并根据这种直接的反馈来更新自身理解的过程。关键在于，一个真正的学习者会被意料之外的结果所**“震惊”（surprised），并必须调整**其对世界的内在模型。而LLM缺乏这种基于真实世界反馈的调整机制。正如萨顿所言：“如果发生了某些事，并非它们所预测的那样，它们也不会因为这件意料之外的事而改变自己。”

萨顿用一段话精辟地概括了这一本质区别：

“我们想要的是一台能从经验中学习的机器，而经验就是你生活中实际发生的事情，你做事，然后看发生了什么……大语言模型学的是别的东西，它们学的是‘这是一个情境，这是一个人的做法’。”

这个区别至关重要，因为它从根本上质疑了LLM是否能真正理解世界。如果一个系统的全部知识都来自于模仿人类的语言模式，而无法被世界的真实反馈所修正，那它究竟是在构建一个关于世界的内在模型，还是仅仅成了一只极其复杂的“鹦鹉”？

2. 根本缺陷：没有“目标”，就没有真正的智能

萨顿对“学习”的严格定义，直接引出了他对LLM范式的第二个、也是更致命的批判：没有目标，就没有真正的智能。他引用了另一位AI先驱约翰·麦卡锡（John McCarthy）的定义：**智能是实现目标的计算能力**。

那么，LLM有目标吗？许多人会说，它的目标是“下一个词元预测”（next-token prediction）。但萨顿一针见血地指出，这并不是一个实质性的目标。它只是一个关于智能体内部状态的指标，而不是一个关乎外部世界的目标。一个系统仅仅因为能准确预测下一个词而“自我感觉良好”，但这并不能称之为智能，因为它没有试图去影响外部世界。

这一缺陷的深层含义是：没有一个真实世界的目标，智能体就失去了判断行为好坏的“基准真相”（ground truth）。它无法在与世界的持续互动中学习和改进。相比之下，强化学习（Reinforcement Learning）的整个框架都建立在一个明确定义目标的“奖励”（reward）信号之上。正是这个奖励信号，驱动着智能体进行真正的学习。

萨顿对此的论断非常有力：

“对我来说，**拥有一个目标是智能的本质**。如果一个东西能够实现目标，它就是智能的……（如果不能），你就不是智能的。”

而我也有类似的观点，作为人类，其最初的“目标”就是生存和繁衍，这也是自私的基因的“自私”的来源。当前的大模型并没有“自私”的动力，也就是Sutton所说的“没有目标", 所以也就“没有真正的智能”。在AI领域，我们也在寻找“自私”的动力，也就是“目标”，来驱动AI的发展。

3. 意外的视角：要理解人类智能，我们应该先研究松鼠

萨顿这种对目标驱动、与环境互动的学习方式的强调，也解释了他为何会提出一个最令人意外的建议。在AI研究领域，普遍观点是我们的目标应是复制人类的独特高级能力，比如制造半导体或登陆月球。然而，萨顿反其道而行之，认为我们应该更多地关注动物与人类的共同点。

这并非一个古怪的奇想，而是他核心理论的逻辑延伸。一只松鼠的整个生命就是他所定义的智能的完美范例：一个持续不断的、由目标驱动的与世界互动来寻找坚果（奖励）和躲避天敌（负奖励）的

过程。萨顿认为，这种学习和智能的基本机制在自然界中是共通的。语言和高级推理等人类独有的能力，只是覆盖在这古老基础之上的薄薄表层。如果我们的AI系统连所有哺乳动物都具备的持续学习能力都没有，又何谈超越人类呢？

他用一句令人难忘的话总结了这个观点：

“**如果我们能理解一只松鼠，我们就几乎走完了通往理解人类智能的全部道路。**”

这个观点之所以强大，是因为它迫使AI领域的研究者们保持谦逊，并提醒我们，在追求那些耀眼的高级认知功能之前，我们可能忽略了智能最根本、最普遍的基础。

4. 历史的教训：“惨痛教训”警示我们警惕LLM的诱惑

2019年，萨顿写下了一篇在AI领域影响深远的文章——《惨痛的教训》（The Bitter Lesson）。其核心思想是：回顾AI历史，那些依赖大规模计算的通用方法（如搜索和学习）最终总是胜过那些试图将人类知识编码进系统的方法。

萨顿认为，LLM是这个“惨痛教训”的一个极具迷惑性的案例。这里存在一个深刻的悖论：LLM一方面拥抱了“惨痛教训”的一半——利用前所未有的大规模计算；但另一方面，它却违背了另一半——它依赖于一个有限的人类知识语料库（互联网文本），而不是一个真正可扩展的、从直接经验中学习的过程。

这正是其“陷阱”所在。依赖人类知识能快速产生令人印象深刻的结果，这种诱惑力是巨大的。然而，萨顿预测，这最终可能是一条死胡同。他坚信，那些能够从与世界直接互动中学习、从而可以从近乎无限的经验数据中获益的系统，最终将超越那些依赖于有限的人类生成语料库的系统。**我们正被LLM的计算规模所吸引，却可能忽略了关于知识来源的更重要教训。这，将成为“惨痛教训”的又一个例证。**

结语：超越模仿，迈向智能新纪元

萨顿的核心信息清晰而深刻：通往真正通用智能的道路，可能不在于建造更大的语言模型，而在于转向一种全新的范式——专注于让智能体通过有目标的、与环境的互动来学习。这个和姚顺雨的"**下半场理论**"(找到正确的问题，和找到恰当的评估方法)有些类似。

更进一步，萨顿为我们描绘了一个关于AI在宇宙中角色的宏大哲学愿景。他认为，我们正在见证一个“伟大的转变”：从“复制者”（replicators）的时代，过渡到“设计”（design）的时代。人类和所有生物都是复制者，我们通过繁衍创造下一代，但并不完全理解其工作原理。而AI则是我们所设计的智能，它的关键特征在于，我们理解其构造原理，并能在此基础上不断改进。这是一种我们能够理解、能够改变、能够构建的智能。

我们正处在这个历史性转变的关口。萨顿并未将其视为威胁，而是将其重塑为人类在宇宙历史中的一个伟大角色——一个新形态智能的骄傲缔造者。他用一句发人深省的话结束了对话，也留给我们无尽的思考：

“**我认为我们应该感到自豪，因为我们正在促成宇宙中这场伟大的转变。**”

附录：通用智能体架构（General Agent Architecture）

Richard Sutton 认为这个架构是解决当前大型语言模型（LLM）困境的根本替代方案和基础模型。Sutton 认为 RL 是**基础人工智能**（basic AI），而 LLM 在概念上是从错误的地方开始的（starting in the wrong place），因此需要一种全新的架构来实现通用智能（AGI）所需的持续、目标驱动的学习。

1. 通用智能体架构的组成

Sutton 提出的通用智能体基础通用模型包含四个核心组成部分：

策略（Policy）： 用于决定在当前情境下智能体**应该做什么**。

价值函数（Value Function）： 用于预测**长期结果**（即长期奖励）。它通过**时序差分学习-TD Learning**习得。价值函数允许智能体将长期目标（如创业十年成功）分解为中间步骤的奖励，通过对长期结果预测的提升来立即强化导致该变化的行动。

感知组件/状态表征（Perception Component / State Representation）： 用于构建智能体对其“现在所处位置”的感知。

世界的转移模型（Transition Model of the World / World Model）： 这是智能体对于如果它采取某个行动**将会发生什么后果**的信念模型。这个模型是通过智能体接收到的**所有感觉信**

息（Sensation）丰富地学习而来的，而不仅仅是从奖励中学习。

2. 如何解决 LLM 的困境

这一通用智能体架构直接针对 Sutton 对 LLM 范式的核心批判点提供了解决方案：

LLM 的困境（Sutton 的批判）	通用智能体架构的解决方案
缺乏目标：LLM 的目标（预测下一个词元）不是实质性的，它们缺乏判断对错或"正确行动"的依据。	**目标驱动的决策**：智能体的核心是拥有目标（即增加奖励）。在 RL 中，"做正确事情的定义是那个能让你获得奖励的事情"。
缺乏经验学习：LLM 学习自训练数据（模仿人类所说的），而非从与世界的实际交互中学习（经验）。	**持续的经验流**：智能的焦点是获取**行动、感觉、奖励**的连续流，并在**正常交互过程中**持续学习，调整行动以增加奖励。
缺乏真正的世界模型：LLM 预测人会说什么，而不是预测**会发生什么**。它们不会对意外事件感到"惊讶"或调整其基本信念。	**包含转移模型**：架构明确包含**世界的转移模型**，该模型预测行动的后果。它通过智能体接收到的**所有感觉信息**（Sensation）进行丰富学习和检验，从而能够理解世界并预测结果。
依赖人类知识（苦涩教训）：LLM 依赖大量人类知识，Sutton 预言这最终会被纯粹从经验和计算中学习的系统超越。	**可扩展的学习机制**：通过价值函数和 TD Learning 等基本、可扩展的方法，智能体能够从经验中构建知识并解决长期目标问题，避免了对人类知识的心理锁定（psychological lock-in）。

总而言之，Sutton 设想的通用智能体架构，通过其四个明确的 RL 核心组成部分，提供了一种范式上的转变，将 AI 研究的重点重新放在**目标设定、持续学习和从经验中构建世界知识**上，以取代当前 LLM 的模仿和监督学习范式。

参考文献

Father of RL thinks LLMs are a dead end

我的公开NotebookLM笔记

分享这篇文章

相关文章推荐

> Agent Lightning

介绍

微软开源的**Agent Lightning**项目，它的核心价值在于为开发者和研究者提供了一个强大的工具，用于**训练和优化 AI Agent（智能代理）**，特别是**几乎不需要修改现有 Agent 代码**就能实现显著的性能提升。

这个项目有以下重要作用：

零代码/低代码训练 AI Agent（核心价值）：

最大亮点： 它允许你使用**强化学习 (Reinforcement Learning, RL)** 等高级优化算法来训练你现有的 AI Agent，而**几乎不需要修改你的 Agent 业务逻辑代码。**这意味着你可以保留你用 LangChain, AutoGen, CrewAI, OpenAI SDK 等框架（甚至裸 Python）编写的 Agent 逻辑，然后让 Agent Lightning 负责优化它的决策过程。

解决痛点： 传统上，将 RL 等技术应用到现有 Agent 框架中需要大量的工程改造和集成工作。Agent Lightning 极大地简化了这个过程。

强大的优化能力：

算法支持： 内置支持**强化学习 (VERL)** 作为核心优化算法，并明确提到支持**自动提示优化 (Automatic**

提供训练基础设施：

> Claude-Code-...

目录

准确性、效率和可靠性i 导到显著提升。

广泛的兼容性和灵活性：

框架无关： 明确支持所有主流 Agent 框架（LangChain、OpenAI Agent SDK, AutoGen, CrewAI）以及纯 Python 实现的 Agent。你可以“即插即用”。

多 Agent 系统优化： 可以在包含多个 Agent 的复杂系统中，**选择性地优化其中一个或几个特定的 Agent**，而不是整个系统，提供了更精细的控制。

Claude Code Router (CCR) 是一款创新的AI模型智能路由工具，它通过拦截 Claude Code 应用对 Anthropic Claude模型的请求，进行多维度分析（如Token数量、用户指令、任务类型），然后依据动态路由规则和配置，将请求智能地导向最合

适的AI模型（来自如Gemini、DeepSeek、本地Ollama模型等不同的模型服务提供商）。CCR的核心机制包括API格式的自动转换与适配、基于Express.js的中间件架构、异步请求处理，以及完善的错误检测、自动降级到兜底模型和潜在的重试策略，旨在提升AI服务调用的效率、灵活性和成本效益。

深入解析 Claude-Code-Router：AI 时代的智能路由中枢

1. 引言：AI 服务智能路由的新范式

在人工智能（AI）技术飞速发展的今天，大语言模型（LLM）

已成为推动各行各业变革的核心引擎。然而，随着模型数量的激增以及它们在能力、性能和成本上的显著差异，如何高效、智能地管理和调度这些模型，以最大化其价值并满足多样化的应用需求，成为了一个亟待解决的关键问题。传统的单一模型服务模式已难以适应日益复杂的应用场景，开发者常常需要在不同模型的 API 之间进行繁琐的切换和适配，这不仅增加了开发成本，也限制了应用的整体性能和灵活性。正是在这样的背景下，**Claude-Code-Router (CCR)** 应运而生，它代表了一种全新的 AI 服务智能路由范式。CCR 通过其精心设计的核心算法与架构，特别是其智能路由决策机制、请求转换与转发策略以及错误处理与降级策略，为多模型的高效协作与按需调度提供了强大的技术支撑。本文将深入探讨 CCR 的这些核心技术，旨在为资深技术专家和架构师提供一个全面而深入的理解，以便更好地评估和应用此类智能路由解决方案，从而在 AI 时代构建更强大、更灵活、更经济的应用系统。

2. Claude-Code-Router 核心机制总览

Claude-Code-Router (CCR) 的核心机制围绕着如何智能地拦截、分析、路由、转换和转发用户请求到最合适的 AI 模型，并将模型的响应有效地返回给用户。这一过程可以概括为一个精细化的处理流水线，确保了请求在整个生命周期中得到高效和准确的处理。CCR 的设计理念在于解耦用户请求与具体模型服务，通过一个中间层来动态管理请求的流向，从而实现模型选择的灵活性、成本的可控性以及服务的鲁棒性。这个中间层，即 CCR 本身，扮演着 AI 服务智能交通枢纽的角色，根据实时的请求特性和预设的策略，将任务分配给最匹配的模型实例。

2.1. 请求拦截与预处理

CCR 的首要步骤是有效地拦截来自客户端（例如 Claude Code 工具）的 API 请求。这是通过一种巧妙的环境变量劫持机制实现的。具体而言，CCR 利用了 Claude Code 工具本身支持通过环境变量

`ANTHROPIC_BASE_URL` 来覆盖其默认 API 端点地址的特性 。通过设置此环境变量，可以将原本直接发送给 Anthropic 官方 API 的请求，重定向到 CCR 本地运行的服务器地址（例如 `http://localhost:3456`）。这种拦截方式无需修改 Claude Code 工具的源代码，实现了对请求流的无侵入式接管，极大地简化了部署和集成过程。一旦请求被成功拦截到 CCR 的本地服务，预处理阶段随即开始。这个阶段主要包括对传入请求的初步校验、日志记录以及为后续的智能路由决策准备必要的上下文信息。例如，CCR 可能会提取请求头中的关键信息，或者对请求体进行初步解析，以确保请求的完整性和有效性，并为后续的分析步骤提供基础数据。

> **Context Engineering**
>
> Context Engineering 是...