Shunyu Yao: The Second Half (下半场)

2025年4月11日 2 分钟阅读

#AI #Shunyu Yao

> *AI的下半场*

Shunyu_Yao，OpenAI的研究员，主要做Agent相关工作。 参与的产品： OpenAI的Operator，Deep Research等 第一作者论文：WebShop，ReAct，Tree of Thoughts，Cognitive Architectures for Language Agents 等等 。 Shunyu_Yao 主页：https://ysymyth.github.io/

“你无法预测未来的点如何连接，但回望过去，一切都清晰可见”-乔布斯。 我的观点：AI的发展也是如此——我们花了几十年专注于算法和模型，打造了一个又一个突破，但现在我们站在转折点上：未来不在于解决已知的问题，而在于定义全新的问题。AI的下半场不是关于技术有多聪明，而是关于我们如何重新设计衡量价值的标尺，创造出真正改变世界的产品。这才是创新的本质——不是跟随，而是引领。

TL,DR

> ***摘要*** *我们正处于人工智能发展的中场阶段。过去几十年，人工智能主要专注于开发新的训练方法和模型，并取得了显著成功。然而，现在的重点正在从解决问题转向定义问题。在这个新的时代，评估比训练更重要，我们需要重新思考如何定义人工智能的目标以及如何衡量真正的进展。*

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #Shunyu Yao

> ***关键点***
>
> *人工智能的第一阶段主要关注开发新的训练方法和模型，取得了诸如DeepBlue、AlphaGo、GPT-4等里程碑式的成果。*
>
> *强化学习（RL）终于能够广泛适用，结合语言和推理的方法，解决了许多复杂任务。*
>
> *RL的关键在于算法、环境和先验知识，而语言预训练提供了强大的先验知识，推动了RL的进步。*
>
> *推理作为一种行动空间的扩展，使得RL能够更好地泛化和适应各种任务。*
>
> *人工智能的第二阶段将重点转向定义问题和重新设计评估方式，而不是仅仅开发新的方法或模型。*
>
> *当前的评估设置与现实世界的应用存在差异，例如自动化和独立任务评估的假设需要被重新审视。*
>
> *第二阶段的目标是开发新的评估设置或任务，以实现真实世界的效用，同时推动人工智能的实用化。*

文章对我的启发

这篇来自姚顺雨（Shunyu Yao）的文章《The Second Half》提出了一个非常精彩且发人深省的观点。作者认为，AI的发展已经走过了以“**发明新方法和新模型**”为核心的“上半场”，现在正式进入了以“**定义问题和衡量价值**”为核心的“下半场”。

根据文章的论述，作者认为在AI的下半场，我们应该采取以下核心应对策略：

核心策略：从根本上重塑评估体系（Fundamentally Re-think Evaluation），将焦点从“提升模型能力”转向“创造真实效用”。

这意味着我们不能再满足于在现有的、固定的基准测试上刷分，而是要主动质疑、打破并创造全新的评估范式（Evaluation Setups），从而倒逼我们去发明真正能够解决现实世界问题的新方法。

具体应对策略的详细分解

作者的逻辑层层递进，可以将其分解为以下几个关键点：

1. 认识到“上半场”游戏的终结

旧规则：上半场的核心玩法是“提出新方法/模型 -> 在现有基准上取得更高分 -> 创造更难的基准 -> 循环”。这个游戏的赢家是像Transformer、AlexNet这样的“方法创新”。

游戏改变者：一个“万能配方”（The Recipe）的出现终结了这场游戏。这个配方由三个关键成分组成：

大规模语言模型预训练（Priors）：作者通过强化学习（RL）的例子雄辩地证明，最重要的部分可能不是RL算法本身，而是通过预训练获得的先验知识。

规模（Scale）：数据和算力的持续扩展。

推理与行动（Reasoning and Acting）：将“思考”作为一种行动，让模型能够规划和泛化。

结果：这个“配方”使得在现有基准上“刷分”变得标准化和工业化。任何一个特定任务上的小幅方法创新，都很容易被下一个“o-series”模型凭借规模优势所碾压。

2. 直面“效用问题”（The Utility Problem）

核心矛盾：尽管AI在各种考试（SAT、Bar）、竞赛（IMO、IOI）和游戏（Go、Chess）中达到了超越人类的水平，但它对真实世界经济和GDP的改变却远未达到预期。

根本原因：我们现有的**评估范式**与**真实世界的工作范式**存在巨大脱节。作者列举了两个典型的例子：

假设1：评估应该是自动化的（Evaluation “should” run automatically）。这导致AI被训练成一个接收长指令、自主运行、返回结果的孤立系统。但在现实中，工作充满了与人的持续互动和沟通。

假设2：评估应该是独立同分布的（Evaluation “should” run i.i.d.）。我们总是独立地测试每一个任务，然后取平均分。但在现实中，任务是连续的，一个软件工程师在同一个代码库里解决的问题越多，他会变得越熟悉、越高效。这需要长期记忆，而现有的i.i.d.评估范式完全忽略了这一点。

3. 开启“下半场”的新游戏

新规则：下半场的新玩法是：

为真实世界的效用，开发新颖的评估范式或任务。

用现有“配方”或增强版的“配方”去解决它们。

循环往复。

思维转变：我们需要从一个纯粹的AI研究员/工程师，转变为一个更接近**产品经理（Product Manager）**的角色。核心问题

从“我们能否训练模型解决X？”变为“**我们应该训练AI去做什么，以及我们如何衡量真正的进展？**”

最终目标：通过创造出那些现有“万能配方”无法轻易解决的、更贴近现实的评估场景，来**迫使**我们去进行真正颠覆性的研究，发明出能够打破现有配方局限的新组件和新方法。

总结来说，作者的呼吁是：不要再沿着旧地图寻找新大陆了。现在最重要的任务，是绘制一张全新的、能够真正通往“实用价值”这个新大陆的地图。这张新地图，就是我们对“评估”的重新定义和根本性创新。

作为全栈工程师和系统架构师，我们能做什么？

这篇文章对我们这些身处一线的工程师来说，有积极的指导意义，我们可以不再仅仅是AI技术的使用者，更可以成为“下半场”游戏规则的定义者。

打造“全栈工作流”评估沙箱 (Full-Stack Workflow Sandbox)

想法: 我们可以利用自己的全栈能力，构建一个高度仿真的“公司内部开发环境”作为评估沙箱。这个沙箱不只是一个代码执行器，它包含：

一个私有的代码仓库（模拟GitLab/GitHub）。

一个项目管理/工单系统（模拟Jira）。

一个内部沟通工具（模拟Slack）。

一个持续集成/持续部署（CI/CD）流水线。

评估任务: 任务不再是“修复这个bug”，而是“认领Jira上的#12345号工单，该工单描述了一个用户反馈的UI显示问题。请在24小时内完成修复、自测、提交代码并通过Code Review，最终部署到预发环境”。

衡量标准: 衡量AI Agent的指标将是：任务完成时间、代码质量、沟通效率（是否需要人类干预）、一次性通过率等。这将直接挑战“i.i.d.”和“自动化”的假设，考验AI的长期记忆和与工具链的协同能力。

设计“从0到1”的产品创造力评估 (Zero-to-One Product Creation Eval)

想法: 创造一个终极评估：给AI一个模糊的商业目标和一个预算。例如，“鹏哥，给你$500的API调用和服务器预算，用两周时间，为独立开发者群体开发并上线一款能提高生产力的SaaS产品。”

评估任务: AI需要自主完成：

市场调研: 使用搜索工具分析竞品和用户痛点。

产品定义: 撰写最小可行产品（MVP）的需求文档。

技术实现: 生成前后端代码、数据库结构，并完成部署。

上线推广: 撰写推广文案，甚至在模拟的社交媒体上发布。

衡量标准: 最终的衡量标准是"产品是否成功上线"、"功能是否满足需求定义"，甚至是"在模拟用户群体中，有多少人愿意'付费'使用"。这直接将评估与最终的"经济效用"挂钩。

构建"人机协作增强因子"评估框架 (Human-AI Collaboration Enhancement Factor)

想法: 彻底放弃评估AI的"单兵作战"能力。我们评估的对象是一个"人类 + AI"的组合。

评估任务: 设计一个对人类专家来说也极具挑战的复杂任务，比如"对一个拥有百万行代码的遗留系统进行现代化重构，并撰写一份完整的技术方案"。

衡量标准: 我们不评估AI本身，而是衡量"增强因子"。例如：

效率增强因子: （专家独立完成时间）/（专家与AI协作完成时间）。

质量增强因子: 由第三方专家组对两个版本的最终成果（独立完成 vs 协作完成）进行盲评打分。

目标: 这种评估范式会迫使AI的发展方向从"替代人类"转向"如何成为最优秀的副驾驶（Copilot）"，从而催生出在交互、意图理解和主动建议方面更强大的模型。

这些新的评估范式本身就是极具价值的AI产品和研究方向。作为全栈工程师，拥有将这些疯狂想法付诸实践的独特优势。 **在AI的下半场，定义问题的人，将比解决问题的人拥有更大的影响力。**

下半场（原文中文翻译-GPT5翻译）

简而言之：我们正处于人工智能发展的中场休息阶段。

几十年来，人工智能的发展主要集中在开发新的训练方法和模型上。这一策略奏效了：从在国际象棋和围棋上击败世界冠军，到在学术能力评估测试（SAT）和律师资格考试中超越大多数人类，再到在国际数学奥林匹克竞赛（IMO）和国际信息学奥林匹克竞赛（IOI）中斩获金牌。在这些载入史册的里程碑背后——深蓝、阿尔

法围棋、GPT-4 以及 o 系列——是人工智能方法的根本创新：搜索、深度强化学习、规模扩展和推理。随着时间的推移，一切都在不断进步。

那么现在到底有什么突然的不同呢？

简而言之：强化学习终于奏效了。更确切地说：强化学习终于能通用了。历经多次重大曲折和一系列里程碑式的进展，我们终于找到了一个可行的方法，能够利用语言和推理解决各种各样的强化学习任务。就在一年前，如果你告诉大多数人工智能研究人员，有一种单一的方法可以应对软件工程、创意写作、国际数学奥林匹克水平的数学、鼠标和键盘操作以及长篇问答——他们可能会嘲笑你的幻想。这些任务每一个都极其困难，许多研究人员整个博士生涯都专注于其中的一个狭窄领域。

然而它还是发生了。

那么接下来会怎样？人工智能的下半场——从现在开始——将把重点从解决问题转向定义问题。在这个新时代，**评估比训练更重要**。我们不再只是问“我们能否训练一个模型来解决 X 问题？”，而是问“我们应该训练人工智能去做什么，以及如何衡量真正的进步？”要在这一下半场中蓬勃发展，我们需要及时转变思维模式和技能组合，或许更接近产品经理的思维。

上半场

要理解上半部分的内容，不妨看看其中的获奖者。到目前为止，你认为最具影响力的 AI 论文有哪些？

我在斯坦福大学的 224N 课程中尝试了测验，答案并不令人意外：Transformer、AlexNet、GPT-3 等。这些论文有什么共同点？它们都提出了一些训练更好模型的根本性突破。但同时，它们也通过在一些基准测试上展示出（显著的）改进成功发表了论文。

不过，有一个潜在的共性：这些“赢家”都是训练方法或模型，而非基准或任务。即便是最具影响力的基准之一 ImageNet，其引用量也还不到 AlexNet 的三分之一。在其他任何地方，方法与基准之间的对比则更为悬殊——例如，Transformer 的主要基准是 WMT'14，其研讨会报告的引用量约为 1300 次，而 Transformer 的引用量则超过了 16 万次。 !image

这说明了**上半场的比赛情况：重点在于构建新的模型和方法**，评估和基准测试是次要的（尽管对于使论文体系运转起来是必要的）。

为什么呢？一个重要原因是，在人工智能发展的前半段，方法比任务更难也更令人兴奋。从零开始创建新的算法或模型架构——想想反向传播算法、卷积网络（如 AlexNet）或 GPT-3 中使用的 Transformer 等突破性成果——需要非凡的洞察力和工程能力。相

比之下，为人工智能定义任务往往感觉更直接：我们只需将人类已经完成的任务（比如翻译、图像识别或国际象棋）转化为基准即可。这并不需要太多洞察力，甚至都不需要多少工程能力。

方法往往比单个任务更具通用性和广泛适用性，因此价值更高。例如，Transformer 架构最终推动了计算机视觉、自然语言处理、强化学习以及许多其他领域的进步——远远超出了它最初证明自身价值的单个数据集（WMT'14 翻译）。一个出色的新方法能够提升许多不同的基准测试，因为它简单且通用，因此其影响往往超越了单个任务。

这款游戏已经运作了数十年，激发了改变世界的创意和突破，这些成果在各个领域不断刷新基准表现。为何游戏规则会有所改变？因为这些创意和突破的累积，已经从质上改变了解决问题的有效方法。

菜谱（The recipe）

秘诀是什么？其成分，不出所料，包括大规模的语言预训练、规模（数据和计算）以及推理和行动的理念。这些听起来可能像是你在旧金山每天都能听到的流行词，但为什么称其为秘诀呢？

我们可以通过强化学习（RL）这一视角来理解这一点，强化学习通常被视为人工智能的"终极目标"——毕竟，从理论上讲，强化学习能确保赢得游戏，而且从经验来看，很难想象有任何超越人类的系统（比如阿尔法围棋）不借助强化学习。

在强化学习中，有三个关键组成部分：**算法、环境和先验知识。**长期以来，强化学习的研究人员主要关注算法（例如 REINFORCE、DQN、TD 学习、策略梯度、PPO、TRPO 等）——即智能体如何学习的智力核心，而将环境和先验知识视为固定不变或次要因素。例如，萨顿和巴托的经典教科书几乎全部围绕算法展开，对环境和先验知识几乎只字未提。

然而，在深度强化学习的时代，从经验来看环境变得非常重要：算法的性能往往高度依赖于其开发和测试所处的环境。如果忽视环境，就有可能构建出一种"最优"算法，但它只在玩具级的设定中表现出色。那么，为何我们不先弄清楚真正想要解决的问题环境，然后再寻找最适合它的算法呢？

这正是 OpenAI 最初的计划。它构建了 gym，这是一个适用于各种游戏的标准强化学习环境，然后是World of Bits 和 Universe项目，试图将互联网或计算机变成一个游戏。这是个不错的计划，不是吗？一旦我们将所有数字世界都变成一个环境，用智能强化学习算法解决它，我们就有了数字通用人工智能。

这是一个不错的计划，但并非完全奏效。OpenAI 在这条道路上取得了巨大进展，利用强化学习解决了《Dota》游戏、机械手等问题。但它从未接近解决计算机使用或网络导航的问题，而且在一个领域中发挥作用的强化学习代理无法转移到另一个领域。显然还缺少了什么。

直到 GPT-2 或 GPT-3 出现之后，人们才发现缺失的部分是先验知识。需要强大的语言预训练将通用常识和语言知识提炼到模型中，然后才能对其进行微调，使其成为网络（WebGPT）或聊天（ChatGPT）代理（并改变世界）。事实证明，**强化学习中最重要的部分可能甚至不是强化学习算法或环境，而是先验知识，而这些先验知识可以通过与强化学习完全无关的方式获得。**

语言预训练为聊天创造了良好的先验条件，但对于控制计算机或玩视频游戏而言效果却不尽如人意。为什么呢？这些领域与互联网文本的分布相去甚远，直接在这些领域上进行有监督微调（SFT）/强化学习（RL）泛化效果不佳。我在 2019 年就注意到了这个问题，当时 GPT-2 刚刚问世，我在其基础上进行了 SFT/RL 来解决基于文本的游戏——CALM是世界上首个通过预训练语言模型构建的智能体。但智能体要爬坡攻克一个游戏需要数百万次的强化学习步骤，而且无法迁移到新游戏中。尽管这正是强化学习的特点，对强化学习研究人员来说也不算奇怪，但我却觉得不可思议，因为我们人类可以轻松玩新游戏，并且零样本表现显著更好。然后我迎来了人生中的第一个顿悟时刻——我们之所以能泛化，是因为我们不只是选择“去2号柜子”“用 1 号钥匙打开3号宝箱”“用剑杀掉地牢”，我们还能选择思考诸如“地牢很危险，我需要一件武器来对抗它。没有看到武器，也许我得在锁着的箱子或宝箱里找一件。3号宝箱在2号柜子里，我先去那里把它打开”之类的事情。

!image

思考或者说推理是一种奇特的行为——它不会直接作用于外部世界，然而推理的空间却是开放且组合上无限的——你可以思考一个词、一个句子、一段完整的文字，或者 10000 个随机的英语单词，但你周围的环境不会立刻发生变化。在经典的强化学习理论中，这是个糟糕的情况，会让决策变得不可能。想象一下，你需要从两个盒子中选一个，其中一个盒子里有 100 万美元，另一个是空的。你预期能赚 50 万美元。现在想象一下，我加入无限个空盒子。你预期将一无所获。但通过将推理纳入任何强化学习环境的动作空间，我们利用语言预训练的先验知识进行泛化，并且能够为不同的决策灵活分配测试时的计算资源。这真是件神奇的事，我在这里没能完全讲清楚，可能需要再写一篇博客专门讲讲。欢迎阅读 ReAct 了解推理在智能体中的原始故事，也可以看看我当时的想法。目前，我的直观解释是：即便你添加了无穷多个空盒子，你一生中在各种游戏中都见过它们，选择这些盒子能让你在任何给定的游戏中更好地选择装有钱的盒子。我的抽象解释是：**语言通过代理人的推理进行概括(language generalizes through reasoning in agents)**。

一旦我们有了合适的强化学习先验知识（语言预训练）和强化学习环境（将语言推理作为行动添加进去），结果发现强化学习算法可能是最简单的一部分。于是我们有了O系列、R1、深度研究、使用计算机的智能体等等更多成果。这是多么讽刺的发展！长期以来，强化学习研究人员对算法的关注远超环境，而且没人关注先验知识——所有强化学习实验基本上都是从零开始。**但经过几十年的弯路，我们才意识到也许我们的优先级应该完全颠倒过来**。

但正如史蒂夫·乔布斯所说：**你无法展望未来时把点点滴滴串联起来；只有在回顾过去时，你才能将它们串联起来。**

下半场（The second half）

这份食谱彻底改变了局面。回顾上半场的情况：

> 我们开发出新颖的训练方法或模型，以在基准测试中不断优化提升。
>
> 我们设定更难的标准，然后不断循环。

这款游戏正在被毁掉，原因在于：

> 该方案实质上已将基准爬坡法标准化和工业化，无需太多新思路。由于该方案具有良好的扩展性和通用性，你针对特定任务的新方法可能使其性能提升 5%，而下一代 O 系列模型则可能在未明确针对该任务的情况下使其性能提升 30%。
>
> 即便我们设定更难的基准，很快（而且会越来越快）它们也会被这种套路解决。我的同事杰森·韦伊制作了一张精美的图表，很好地展现了这一趋势：

!image]

那么下半场还有什么可玩的呢？如果不再需要新颖的方法，而更难的基准测试也会越来越快地被攻克，那我们该怎么办？

我认为我们应当从根本上重新思考评估方式。这不仅意味着要制定新的、更严格的基准，还意味着要从根本上质疑现有的**评估体系**(evaluation setups)并创建新的体系，从而迫使我们发明超越现有工作方法的新方法。这很难，因为人类有惯性，很少质疑基本假设——你只是想当然地接受它们，却没意识到它们只是假设，而非定律。

要解释惯性，假设你在 2021 年发明了一种基于人类考试的最成功的评估方法之一。这是一个极其大胆的想法，但三年后它已趋于饱和。你会怎么做？很可能会设计出更难的考试。或者假设你只是解决简单的编程任务。你会怎么做？很可能会寻找更难的编程任务来解决，直到达到国际信息学奥林匹克竞赛金牌水平。

惯性是自然存在的，但问题在于，人工智能已经在国际象棋和围棋上击败了世界冠军，在 SAT 和律师资格考试中超越了大多数人类，在国际信息学奥林匹克竞赛和国际数学奥林匹克竞赛中达到了金牌水平。但世界似乎并未发生太大变化，至少从经济和国内生产总值（GDP）的角度来看是这样。

我将此称为**效用问题(utility problem)**，并认为这是人工智能面临的最重要的问题。

或许我们很快就能解决效用问题，或许不能。不管怎样，这个问题的根本原因可能看似简单：**我们的评估设置在很多基本方面都与现实世界的设置不同**。举两个例子：

> **评估“应当”自动运行**，所以通常情况下，智能体接收任务输入，自主完成任务，然后获得任务奖励。但在现实中，智能体在整个任务过程中必须与人类互动——你不会只是给客服发一条超级长的消息，等上 10 分钟，然后就指望得到一个详尽的回复来解决所有问题。质疑这种设置，新的基准测试被发明出来，要么让真实的人类参与其中（例如 Chatbot 领域），要么进行用户模拟（例如 tau-bench）。 !image
>
> **评估“应该”独立同分布地进行**。如果你有一个包含 500 个任务的测试集，那么你应独立运行每个任务，对任务指标取平均值，从而得出总体指标。但在实际中，任务是按顺序而非并行解决的。谷歌的一名软件工程师在对代码库越来越熟悉的情况下，解决 google3 问题的能力会逐渐增强，但一个软件工程师代理在解决同一代码库中的许多问题时却不会获得这种熟悉度。显然，我们需要长期记忆方法（而且确实存在），但学术界没有合适的基准来证明这种需求，甚至没有足够的勇气去质疑作为机器学习基础的独立同分布假设。

这些假设“一直”都是如此，在人工智能发展的前半段，基于这些假设来制定基准是没问题的，**因为当智能水平较低时，提升智能通常会提高效用**。但如今，在这些假设下，通用的方案已无法保证奏效。所以，要玩好人工智能发展的后半段这场新游戏，方法是：

> 我们开发新颖的评估方案或任务以用于实际应用。
>
> 我们用配方来解决这些问题，或者用新颖的成分来丰富配方。
>
> 如此循环往复。

这款游戏很难，因为它很陌生。但它也很令人兴奋。在前半段，玩家要解决视频游戏和考试，而在后半段，玩家则通过运用智慧打造有用的产品来创建价值数十亿甚至数万亿美元的公司。前半段充斥着渐进的方法和模型，而后半段则在一定程度上对它们进行筛选。一般的套路会轻易击垮你的渐进方法，除非你提出新的假设来打破套路。这样你才能进行真正具有变革性的研究。

欢迎来到下半场！

致谢 这篇博客文章基于我在斯坦福大学 224N 课程和哥伦比亚大学所做的演讲。我借助了 OpenAI 的深度研究来阅读我的幻灯片并撰写初稿。

写于 2025 年 4 月 10 日

来源

https://ysymyth.github.io/The-Second-Half/

CAMEL Tools

CAMEL工具包是一个模块化框架，旨在通过统一接口扩展AI智能体的能力，使其能够连接外部服务、数据源和计算工具。它提供了多种工具包，涵盖搜索、学术、社交媒体、数据分析、媒体处理、开发、金融和生产力等领域，帮助开发者加速开发、提升可靠性并简化API集成。

CAMEL工具包通过一致的API设计（基于BaseToolkit类）和模型上下文协议（MCP）标准化了工具使用，简化了学习和实施过程。

工具包解决了API集成开销、不一致的接口、网络和错误处理以及维护问题。

主要工具包包括：

网络和搜索工具包：支持多种搜索引擎和知识库，提供

实时数据访问。

学术和研究工具包：如arXiv、Google Scholar、PubMed等，专注于学术文献检索和分析。

社交媒体和通信工具包：如Twitter、Reddit、LinkedIn等，支持社交媒体数据分析和交互。

数据分析和计算工具包：如数学、SymPy、NetworkX等，支持数学运算、网络分析和数据处理。

媒体处理工具包：如DALL-E、音频分析、视频分析等，用于图像、音频和视频内容的生成和分析。

开发和编码工具包：如GitHub、终端、代码执行工具包等，支持开发者任务自动化。

金融和商业工具包：如Stripe、OpenBB等，支持支付处理和金融数据分析。

生产力和集成工具包：如MCP、Notion、Excel等，支持项目管理、文档处理和跨平台集成。

CAMEL工具包的优势包括：加速开发、一致接口、可组合性、可靠性与未来兼容性。

不同工具包适用于不同场景，如信息获取、业务优化、创意生成、开发辅助和复杂AI系统。

CAMEL框架通过模块化设计支持工具包的轻松更新和扩展，满足不断变化的市场需求。

1. 网络和搜

索类工具包

工具包名称	主要功能	适用场景
搜索工具包	• Google、Bing、DuckDuckGo等搜索引擎集成 • Tavily、Linkup专业搜索 • Wikipedia、Wolfram Alpha知识库访问	• 事实查询 • 最新信息获取 • 研究助手开发
浏览器工具包	• 网页导航 • 内容提取 • 表单填写 • 会话管理	• 网站数据抓取 • 表单自动化 • 电商助手开发
天气工具包	• 全球天气数据获取 • 天气预报 • 历史记录查询	• 旅行规划 • 物流

工具包名称	主要功能	适用场景
		路线优化·环境感知服务

2. 学术和研

究类工具包

工具包名称	主要功能	适用场景
Arxiv工具包	• 科学论文搜索 • 按关键词/作者/类别检索	• 研究助手 • 预印本跟踪 • 文献综述
Google Scholar工具包	• 学术出版物检索 • 引用信息分析 • 作者资料查询	• 跨出版商搜索 • 文献计量分析 • 研究影响力追踪
PubMed工具包	• 生物医学文献访	• 医学研究 • 临床

工具包名称	主要功能	适用场景
	问 • 临床研究数据库检索	决策支持 • 制药研究
Semantic Scholar 工具包	• 语义相关性搜索 • AI 驱动的文献分析	• 语义分析 • 跨学科研究 • 趋势识别

3. 社交媒体和通信类工

具包

工具包名称	主要功能	适用场景
Twitter工具包	• 推文检索 • 话题跟踪 • 个人资料分析	• 社媒监控 • 品牌声誉管理 • 趋势分析
Reddit工具包	• 帖子检索 • 评论分析 • 讨论跟踪	• 内容聚合 • 情感分析 • 趋势发现
LinkedIn工具包	• 专业资料检索 • 公司数据分析 • 职	• 招聘助手 • 职业发展 • 商业智能

工具包名称	主要功能	适用场景
	位信息获取	
Slack工具包	• 消息发送 • 频道管理 • 对话历史记录	• 工作效率工具 • 团队协作 • 工作流集成
WhatsApp工具包	• 消息收发 • 联系人管理 • 聊天记录访问	• 客服机器人 • 预约提醒 • 电商通讯

4. 数据分析和计算类工

具包

工具包名称	主要功能	适用场景
数学工具包	• 基础到高级运算 • 单位转换	• 金融计算 • 工程计算 • 数据科学
SymPy工具包	• 符号数学运算 • 微积分计算 • 矩阵处理	• 高级数学教育 • 工程研究 • 定理证明
NetworkX工具包	• 图分析 • 网络结构创建 • 图算法	• 社交网络分析 • 路由优化 • 推荐

工具包名称	主要功能	适用场景
	实现	系统
Data Commons 工具包	• 公共数据访问 • 统计分析 • 人口统计	• 政策分析 • 社会经济研究 • 公共卫生

5. 媒体处理

类工具包

工具包名称	主要功能	适用场景
DALL-E工具包	• 文本生成图像 • 图像修改 • 风格控制	• 创意设计 • 营销原型 • 概念可视化
音频分析工具包	• 语音识别 • 声音分类 • 语音分析	• 语音助手 • 内容审核 • 音乐推荐
视频分析工具包	• 对象检测 • 场景分析 • 动作识别	• 内容管理 • 安全监控 • 运动分析

工具包名称	主要功能	适用场景
图像分析工具包	• 对象检测 • 图像分类 • OCR 识别	• 文档扫描 • 内容过滤 • 图像搜索
视频下载工具包	• 视频检索 • 格式转换 • 元数据提取	• 内容存档 • 教育培训 • 媒体分析

6. 开发和编码类工具包

工具包名称	主要功能	适用场景
GitHub工具包	• 代码仓库交互 • 提交管理 • 问题跟踪	• 编码助手 • 代码分析 • 项目管理
终端工具包	• 系统命令执行 • 脚本运行 • Shell交互	• DevOps任务 • 环境配置 • 系统管理
代码执行工具包	• 多语言代码运行 • 沙盒环境支持	• 编程教学 • 代码测试 • 算法实验
文件写入工具包	• 文件创建修改 • 权限管理	• 文档生成 • 配置管理 • 内容自动化

7. 金融和商

业类工具包

工具包名称	主要功能	适用场景
Stripe工具包	• 支付处理 • 订阅管理 • 客户数据管理	• 电商支付 • 订阅业务 • 财务分析
OpenBB工具包	• 金融数据分析 • 市场可视化	• 投资咨询 • 风险评估 • 投资组合追踪

工具包名称	主要功能	适用场景
MinerU 工具包	• 文档处理 • OCR 识别 • 表格检测	• 内容提取 • 公式识别 • 数据结构化
Dappier 工具包	• 实时数据访问 • AI 推荐	• 信息检索 • 内容聚合 • 数据分析

8. 生产力和集成类工具

包

工具包名称	主要功能	适用场景
MCP工具包	• 多服务器连接管理 • 工具生命周期控制	• 大规模AI系统 • 工作流分解 • 协作环境
Notion工具包	• 页面管理 • 数据库处理 • 内容上传	• 知识库构建 • 项目管理 • 团队协作
Excel工具包	• 电子表格处理 • 格式保留	• 数据提取 • 文档转换 • 数据

工具包名称	主要功能	适用场景
		分析
Zapier工具包	• 自然语言命令 • 工作流自动化	• 流程自动化 • 服务集成 • 任务执行
Open API工具包	• API集成 • 请求处理	• 多API管理 • 服务代理 • API测试
AskNews工具包	• 新闻聚合 • 情感分析	• 新闻摘要 • 媒体监控 • 趋势检测
Meshy工具包	• 3D模	• 产品

工具包名称	主要功能	适用场景
	型生成• 模型编辑	设计• 建筑可视化• 游戏内容
Human 工具包	• 用户输入管理• 反馈收集	• 人机协作• 模型优化• 决策验证

这些工具包展现了CAMEL框架强大的生态系统，能够满足从基础开发到高级AI应用的各种需求。每个工具包都经过精心设计，既可以独立使用，也可以组合使用以构建更复杂的应用。