

# 从社区数据再看大模型开发生态

在一年前和我司 CTO 探讨开源和技术势态的时候，我们碰撞出来一个观点——“作为一个开放、领先的科技公司，我们应该可以主动利用社区数据，形成自己对技术趋势的洞察”。由此，我们开始尝试基于开源社区的行为数据，对技术趋势进行分析。并且，“来自于社区，回馈到社区”，我们不仅会把这个分析的到的全景图和结论分享给社区，也会把过程中的数据分享出来。

在上半年的“527 蚂蚁技术日”，我们发布的全景和趋势得到了很多关注和肯定，更重要的是，我们的同事和朋友们都在技术和架构判断，技术选型，兼容性取舍，甚至是商业拓展选择上都有参考我们的全景图。当然，发布后的这三个月，我们也收到很多意见、建议和疑问，同时社区也发生了许多变化，于是，在外滩大会上，我们下半年的发布也如期而至了。

—— 王旭，蚂蚁开源技术委员会副主席

## 100 天前，我们发布了一份全景图

3 个多月前，在一年一度的「527 蚂蚁技术日」上，我们发布了一份大模型开发生态下的开源项目全景图，和一份对生态趋势的洞察报告。这是一份利用开源社区中的数据制作的全景图，在寻找这个领域生态下究竟存在哪些项目，以及应该用怎样的评价标准来判断项目的核心程度与热门程度这两个关键步骤中，都使用到了开发者们在开源平台 GitHub 上产生的协作数据。



(图源：527 蚂蚁技术日，蚂蚁开源技术委员会副主席 – 王旭)

发布之后，收到了很多来自社区的点赞和非常有价值的反馈。很感谢大家的鼓励，认可这是一件有价值的工作，并在过去的三个月陆续在各种渠道上分享、引用甚至进行二次的创作。发布之初，我们收到了一些灵魂发问：

- 为什么是蚂蚁来做这件事情？上面偷偷塞了多少个蚂蚁的开源项目？

当发布一张全景图时，不可避免的，很容易受到关注的视角是它够不够“全”，我们最常听到的提问是“xxx项目怎么没有在上面”。而由蚂蚁这样一家技术公司来发布这件事，还容易被关注它究竟是否“客观”。首先，如之前所述，蚂蚁做这件事情的初衷，并不是像 CNCF 基金会的全景图那样，容纳尽可能多的生态项目，而是希望了解生态中在主轴线或架构核心位置的项目究竟是哪些，在服务于内部决策的参考的同时，也共享给社区。因此我们使用了 OpenRank 这样的开源指标，划定了一个参考标准，对于大多数领域，如果符合了这个评价标准，才会被放在全景图上。当然，社区数据毕竟难以做到全面，依旧会有非常有价值的、社区活跃的项目被遗漏，所以我们也 GitHub 上开启了一条 Issue 作为反馈的公共入口：

- <https://github.com/antgroup/lm-oss-landscape>

还有一些注重细节的同学，提醒我们有一些项目虽然代码开放了，但是使用的许可证没有经过 OSI 认证，严格上讲不能算作开源项目，在这次的洞察中，我们也对这种情况做了剖析。

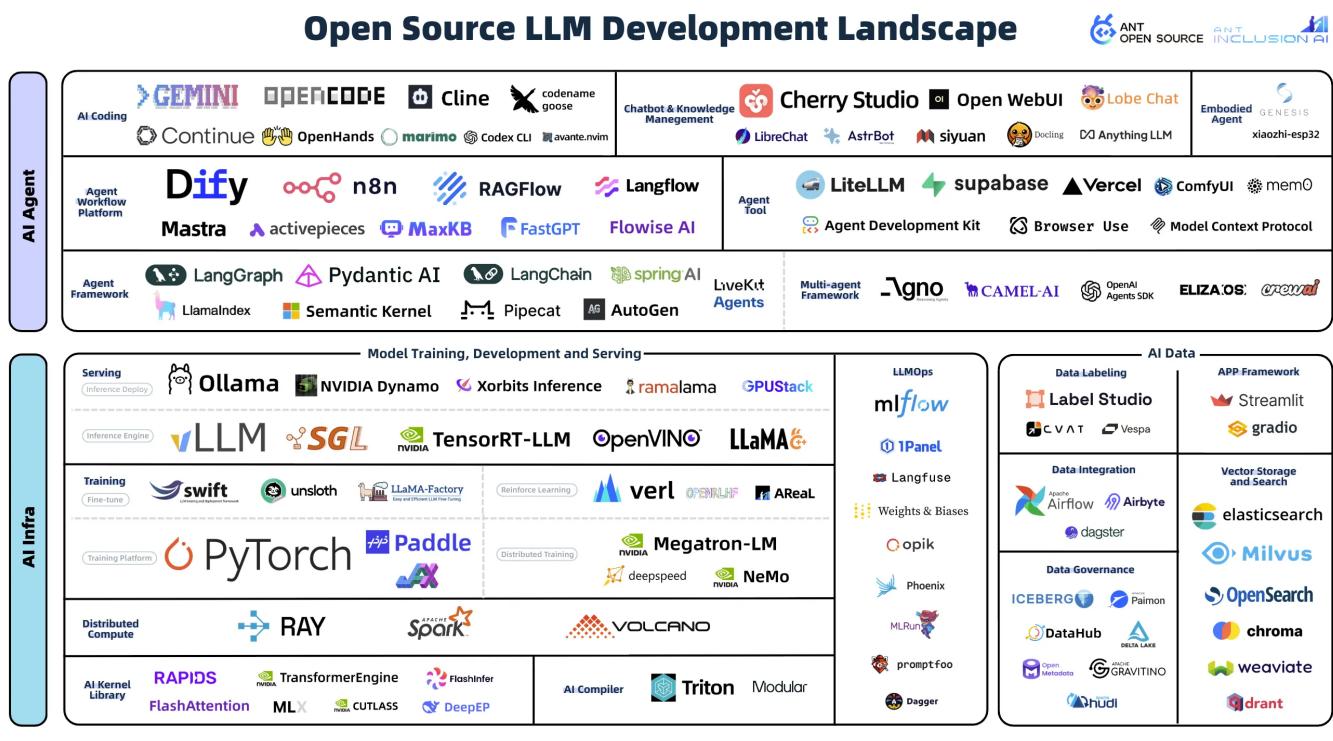
大模型 AI 生态是当下毫无疑问备受瞩目的热点，但像操作系统、云原生和一些相对小众的技术领域仍然是开源和技术世界里不可或缺的组成部分，在这些领域也存在着闪闪发光的项目们，构筑着我们数字世界的基石。在这些领域的开发者们，关心能否复刻我们制作全景图时的方法论和工具，在他们所关心的技术领域，也能够复现出由其中的高光项目们所组成的全景图。

在这些反馈的基础上，我们对看生态全景的方法进行了更新，并且将如何标注项目、消费数据的细节更新在了 GitHub 上供大家参考。第一版里，我们通过一些已知的种子项目（PyTorch、vLLM、LangChain），基于开发者的协作关系多跳搜索和它们紧密关联的开源项目，这种方式受到选取的种子项目、每跳返回的项目数量等因素影响，得到的结果具有很大程度的随机性。而全景图使用的评价方法 OpenRank 本身就是一种基于社区协作关联关系，计算生态中所有项目的相对影响力的方法。因此在这一次，我们直接拉取了当月 GitHub 全域项目的 OpenRank 排名，根据描述和标签来从上往下标注出其中属于大模型生态的项目，再逐步收敛。果然，这个过程中发现了更多之前未发现的、热度和活跃度都相当高的项目们，让我们可以自信的将参考标准提高至了当月 OpenRank 达到 50 这个水平。

注：OpenRank 是一种社区导向的算法，在直接面向开发者、或者开发者群体基数本身就比较大领域，值往往是更高的，但对于更底层、开发者基数更小的项目，例如算子库、编译器等，则相对难以体现出项目的价值，因此，我们不可避免地仍旧引入了一些人为判断的成分。

# 再看生态与趋势

## 大模型开发生态全景



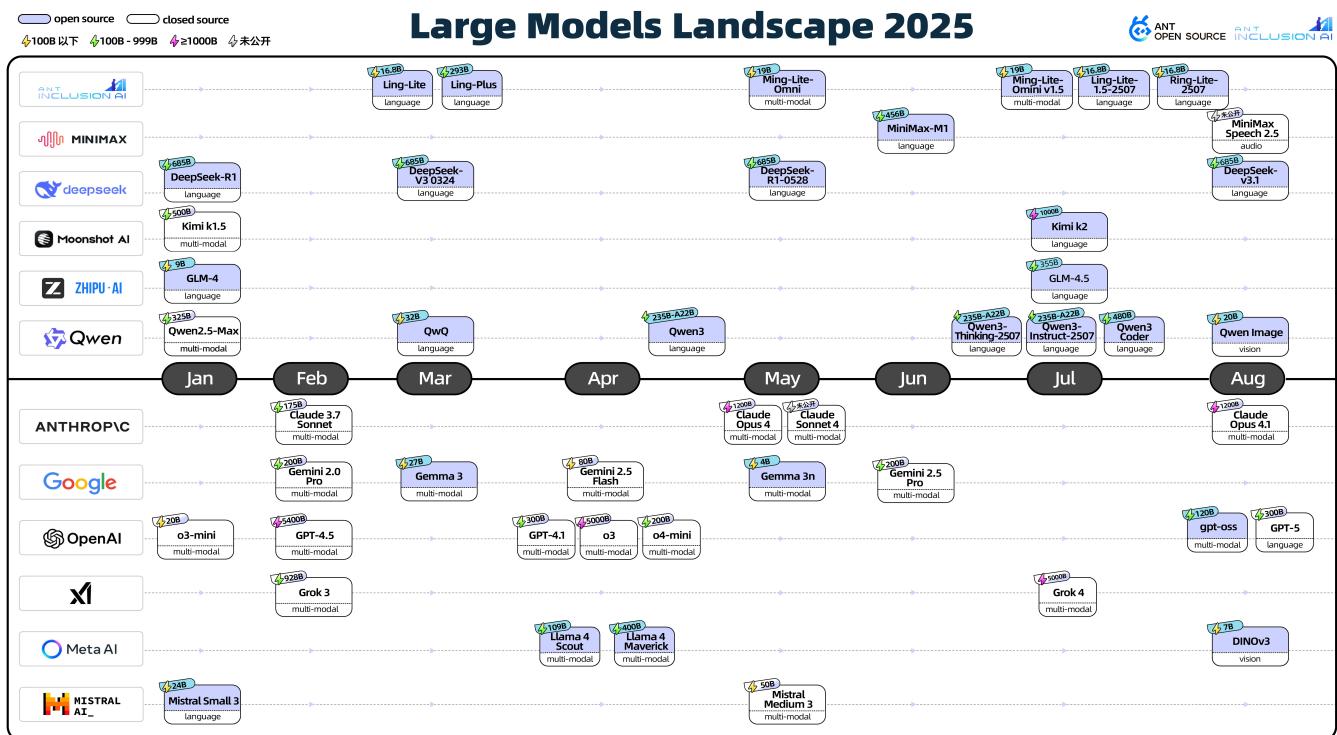
<https://antoss-landscape.my.canva.site/>

因为评价标准的提高，在这次发布的全景图中，收录的项目收敛为了 114 个，覆盖了 22 个技术领域分类。其中，有 39 个项目是这次新进的，占据当前整体版面的 35%。而第一版中的 60 个项目被拿掉了，这背后最主要的原因是项目达不到 **OpenRank 大于 50** 这个新的标准了，而其中有不少从趋势来看，也确实已经在步入 AI 墓园的路上，后面我们会详细展开。也有部分项目，典型的如 ONNXRuntime，由于主要面向于传统机器学习的训练和推理，在大模型领域并没有很紧密的结合而被拿掉。

算上那些被拿掉的在大模型开发生态的项目们，这些项目整体从创建至今的年龄分布在 30 个月这个中位数，也就是两年半。他们年轻的程度正应和着这个领域迭代的速度，高达 62% 的项目都是在“GPT”时刻（2022 年 10 月）之后发布的，而其中有 12 个项目甚至是在今年才新近发起的。在此如此崭新的基础上，这些项目获得的关注度却是上一个时代的开源项目们难以企及的：它们平均获得的星标数量高达近 3 万个。

这些项目吸引了全球 366,521 位开发者的参与。在能够统计到位置信息的开发者中，约 24% 来自美国，18% 来自中国，其次是印度（8%）、德国（6%）和英国（5%）。无论是大模型的研发还是围绕着模型的开源开发生态，美国和中国都扮演着主导角色，这一格局也许会进一步影响全球技术的演进与合作。

## 大模型 2025 发布时间线



在开源开发生态之外，模型也在进行高频的发布。虽然目前还没有很好的数据渠道来帮助我们理解大模型社区，但毕竟它们处在注意力焦点的中心，因此，这里也梳理了一些我们自身也比较关注的模型厂商在 2025 年发布的大模型的时间线，这其中包含了开放参数的模型，也有闭源的模型。我们也标注了每个模型的参数、模态这些关键信息，来一定程度上帮助理解当下各家厂商的白热化竞争究竟是在哪些方向上较劲。

- 中美开源与闭源的路线分化：**从这张图上，能够更直观的感受到中国开源大模型的百花齐放，而国外的顶尖模型厂商依旧走的是闭源路线。在早期几乎凭借一己之力对抗模型封闭生态的 Meta（也使得不少在全景图上的开源项目名字中都带有“llama”的元素），似乎也正在考虑向闭源转型：7月 31 日，Meta CEO 扎克伯格发布了一封名为「Personal Superintelligence」的公开信，表达对超级智能个体的野心的同时，还说了这样一句话：“我们会更谨慎的选择开源什么”。Meta 在今年的

AI 战略发展确实不尽人意，各大评测榜单上已经看不到 Llama 系列的身影，而四月份发布的 Llama 4 也陷入了“效果差”甚至是“造假”的争议。

- **MoE 架构下模型参数的规模化发展：**今年发布的 DeepSeek、Qwen、Kimi 等旗舰模型全面采用了专家混合 (Mixture of Experts, MoE) 这种神经网络架构思想，它最朴素的原理为“稀疏激活”：虽然模型总参数可以非常庞大，但每次推理时只用其中很小一部分。在这种架构下，我们看到了 K2、Claude Opus、o3 等达到了万亿参数规模的庞大模型在今年陆续发布。参数规模的增加能够有效提升模型在任务上的表现，但同时也对训练和推理时的计算与内存提出了进一步的挑战。
- **通过强化学习提升模型 Reasoning 能力：**DeepSeek R1 通过将强化学习后训的过程与大规模预训练结合，显著提升了模型性能，在自动化推理、复杂决策和知识推断等任务上，比传统的 LLM 提高了多个维度的能力，Reasoning 能力也成为了今年重磅模型在发布时的时尚单品。由于模型在推理时普遍需要更久的时间和更多的 token，Qwen、Calude、Gemini 等系列模型也逐步整合了“混合推理”的能力：如同人类大脑有快速反应和深度思考两种模式，用户也可以基于需求场景，让模型在不同模式下给出反应。
- **多模态模型走向主流：**当前市面上的多模态模型支持的能力以语言、图像和语音的交互为主，也有一些垂类的视觉模型和语音模型在今年发布，而在开发生态中，我们也发现了围绕着语音模态的丰富工具链，如 Pipecat、LiveKit Agents 和 CosyVoice。在 2024 年年初，OpenAI 的发布的 Sora 演示视频惊艳世界，有关世界模型和通用人工智能似乎已经不再停留于畅想，而站在 2025 这个时间节点，无论是视频模态的成熟还是 AGI 的成功，都仍旧有一段路要走。
- **主观和客观两种模型评价方式：**对模型的评价和排名，整体可以分为两种模式：
  - 基于人类主观投票的评测。代表平台：[Design Arena](#), [LMArena](#)
  - 基于客观标准答案的评测。下表梳理了最近两个月新模型发布时主要提及的性能对比评测集，可以作为当下最顶尖也最前沿的评测集的代表：

当下的主流评测集	面向领域	被哪些最新发布的模型使用
AIME 2025	Math	DeepSeek-V3.1, GPT-5, Claude Opus 4.1, Qwen3-2507, Kimi K2, Grok 4
GPQA Diamond	Math	GPT-5, Claude Opus 4.1, Kimi K2 (GPQA: Grok 4, Qwen3-2507, GLM-4.5)
LiveCode Bench v6	Math	Qwen3-2507, Kimi K2
SWE-bench verified	Coding	DeepSeek-V3.1, GPT-5, Claude Opus 4.1, Kimi K2, GLM 4.5
Terminal-Bench	Coding	DeepSeek-V3.1, Claude Opus 4.1, GLM 4.5
BrowseComp	Agentic	DeepSeek-V3.1, GPT-5, GLM 4.5



3. Data: 79 次
4. Learning: 44 次
5. Search: 36 次
6. Model: 36 次
7. OpenAI: 35 次
8. Framework: 32 次
9. Python: 30 次
10. MCP: 29 次

## OpenRank Top 10

#	Project	Ran k	Domain	Open Rank	Star s	OpenRank Trend	Langu age	Created	Initiator	License
1	<a href="#">PyTorch</a>	-	Training Platform	859	920 39		Python	2016–08–13	Meta	BSD-3 Clause
2	<a href="#">vLLM</a>	-	Inference Engine	637	5391 2		Python	2023–02–09	Berkeley	Apache 2.0
3	<a href="#">Gemini CLI</a>	new	AI Coding	391	668 81		TypeScript	2025–04–17	Google	Apache 2.0
4	<a href="#">Dify</a>	-1	Agent Platform	379	1096 74		TypeScript	2023–04–12	Dify.AI	 Open Source License
5	<a href="#">SGLang</a>	+3	Inference Engine	352	1655 5		Python	2024–01–08	Berkeley	Apache 2.0
6	<a href="#">Airflow</a>	+1	Data Integration	323	4137 1		Python	2015–04–13	Airbnb	Apache 2.0

7	<a href="#">cherry-studio</a>	new	Chatbot	313	309 76		TypeScript	2024-05-24	Shanghai Qianhui Tech	⚠️ User-Segment Dual License
8	<a href="#">TensorRT-LLM</a>	new	Inference Engine	295	1121 9		C++	2023-08-16	NVIDIA	Apache 2.0
9	<a href="#">n8n</a>	+6	Agent Platform	281	1266 59		TypeScript	2019-06-22	n8n	⚠️ Sustainable Use License
10	<a href="#">Ray</a>	+2	Distributed Compute	274	383 02		Python	2016-10-25	Anyscale	Apache 2.0

头部这 10 个项目，代表了当下大模型开发生态里最活跃、最具代表性的社区力量。它们几乎覆盖了模型生态的完整链路：从底层的算力和框架 PyTorch、Ray，模型训练的数据处理管线 Airflow，模型服务的性能基座 vLLM、SGLang、TensorRT-LLM，到 Agent 应用调度平台 Dify、n8n，直接面向开发者与终端用户的 Gemini CLI、Cherry Studio。从编程语言来看，Python 主导基础设施，TypeScript 统治应用层，成为支撑整个生态体系的核心语言。而从背后的发起力量来看，我们看到了来自学术界的创新迸发出的高影响力：vLLM、SGLang、Ray 都生长于 Ion Stoica 执掌下的伯克利实验室；Meta、Google、NVIDIA 这些大厂掌控或布局在一些关键节点之上，但在靠近应用层的位置，Dify、Cherry Studio 这样的独立团队也能够迅速创新，通过提供用户友好的工具，形成快速增长点。

## 如何定义这个时代的开源？

熟悉围绕开源许可证的一些前尘往事的开源老人，在看到刚刚这 10 个顶尖的项目所采用的许可证时，也许心中已经警铃大作。是的，虽然多数大模型开发生态的项目仍然采用的是 Apache 2.0 或 MIT 宽松许可，但仍然有不少值得关注的特别案例：

- **Dify 的 Open Source License**。这是 Dify 基于 Apache 2.0 许可的文本做了修改，增加了两个附加条款：

- a. 限制未经许可的多租户环境运营；  
b. 使用Dify前端时，不得移除或修改 LOGO 和版权信息
- **n8n 的 Sustainable Use License。** 这是 n8n 基于 fair-code 主张，新提出的一种许可，在允许免费使用、修改、分发的基础上，做了三点限制：
  - a. 仅限于在企业内部，或者非商业、个人用途下使用或修改；
  - b. 在分发时，必须基于非商业目的免费提供；
  - c. 不能更改软件中的许可、版权或作者信息
- **Cherry Studio 的 User-Segmented Dual Licensing。** Cherry Studio 根据用户所在组织的规模做分段，提出了一种双许可限制，不同规模组织下的用户使用不同的许可：
  - a. 如果是个人用户或者所在组织是 10 人及以下，采用 AGPLv3，这也是一种 copyleft 协议，用户可以免费使用，但如果做了修改和分发，必须同样开源并提供完整的源代码；
  - b. 超过 10 人的情况，则需要联系 Cherry Studio 的团队进行商业授权。

可以看出，上述许可证的条款多半是出于保护商业利益的考虑，由于带有对部分用户的“歧视”属性，自然难以获得 OSI 的批准。从开源原教旨主义的角度来看，它们甚至未必算得上真正的开源项目。

在当下，「开源」的定义愈发模糊：不仅“开源大模型”与“开放权重大模型”之间存在诸多争议，传统软件的开源也仿佛在雾里看花。与此同时，GitHub 不再只是单纯的代码托管、协作和分发平台，而是成为这一时代的运营阵地：许多连源代码都闭源的产品（如 Cursor、Claude-Code 等）依旧在 GitHub 上占有一席之地，让看客们常常拥有一种它们也是开源项目的错觉。这些仓库无一例外拥有一骑绝尘的 star 数量，但它承担的真正功能也许只是作为厂商收集用户反馈的入口。

其他非 OSI 批准许可：

- Elastic: Elasticsearch 和 Airbyte

OSI 批准的其他值得注意的被使用许可：

- GPLv3 (copyleft 许可，对分发作了严格限制) : 1Panel/MaxKB, 1Panel/1Panel, Comfy UI
- BSD3-Claude “New” or “Revised” License (非常宽松的许可，可以放心使用) : PyTorch, Weaviate, Flash-attention

## 技术领域的趋势

从今年的技术领域发展趋势来看，AI Coding、Model Serving 和 LLM Ops 整体处于增长的态势，尤其是 AI Coding 的增长斜率在近两个月还在持续攀升，再次印证了 AI 研发提效是 2025 年真正被验证和落地的应用场景；Agent Framework 和 AI Data 是下跌比较明显的两个领域，Agent 开发框架的下跌和

曾经在头部的 LangChain、LlamaIndex、AutoGen 等项目在社区投入上的显著收缩有很大关系，而 AI Data 在向量存储、数据集成及数据治理等维度上，也表现出在平稳中逐步下降的趋势。



## 边缘地带的项目们

如下是本次没有出现在全景图上，但是依旧被认为是很有潜力的开源项目们，我们会持续保持关注。继续加油！

Project	Open Rank	Star	OpenRank Trend	Language	Created	Comment
<a href="#">InvokeAI</a>	48	2563		TypeScript	2022-08-17	为 Stable Diffusion 模型提供的 WebUI 创作引擎。类似的项目还有 ComfyUI、stable-diffusion-webui，都拥有更加可观的 Star 数量和更为陡峭的 OpenRank 下降曲线。
<a href="#">onyx</a>	46	1325		Python	2023-04-27	一种锚定了团队协作场景的 Chatbot，基于 GenAI 的 Teams 聊天工具 – 把你团队的专有知识喂给大模型。
<a href="#">deer-flow</a>	37	1595		Python	2025-05-07	字节推出的 Deep Research 框架，在模型之上集成了 Web 搜索，数据抓取和脚本执行的能力，一经推出即受到关注，但近两个月维护度下降，社区数据逐渐跌落。
<a href="#">Mooncake</a>	36	3704		C++	2024-06-25	清华大学 KVCache.AI 团队提出的模型服务平台，虽然关注度和社区指标都不算高，但能够看到明显的攀升走势。
<a href="#">KTransformers</a>	34	1478		Python	2024-07-26	同为 KVCache.AI 提出的推理优化框架，在今年 2 月实现了本地单机部署千亿参数满血版的 DeepSeek 模型之后迅速爆火，随后持续回落。
<a href="#">CosyVoice</a>	32	1554		Python	2024-07-03	多语言语音生成大模型，模型开源的同时，也开源了推理，训练和部署的全栈工具链。近几个月数据稍见颓势，还需继续观望。

A2A	29	1882	5	TypeScript	2025-03-25	A2A 协议在今年 MCP 最火热的时候由 Google 提出，并随后在 6 月份官宣捐赠给 Linux 基金会。作为大厂占据生态位的战略布局，A2A 无论是社区化还是被接纳的程度，都需要等待时间验证。
-----	----	------	---	------------	------------	---

## 大模型生态下全球开发者分布画像

两次发布的全景图涉及到的一共 170 多个开源项目中，在其中有过 Issue 或 PR 相关行为的 GitHub 账号高达 36 万，这个数字一定程度上体现了当下大模型生态的开发者规模。我们识别到其中 124,351 位在个人页面填写了可以被正确解析位置信息的开发者，并统计了他们的国家分布和对应的在大模型开发生态中的贡献度分布。图和表中展示了头部国家的**开发者贡献度总和、整体贡献度占比和识别到的开发者数量**，其中，将开发者数量乘以三的话，可以大致认为是估算出的该国家大模型生态开发者的总量。

总体来看，中美引领了 AI 领域的开源贡献。美国以 37.4% 的贡献领先，中国以 18.7% 位居第二，这两个国家的贡献总比例达到 55% 以上，而排名第三的德国已降低至 6.5%。

注：开发者贡献度也使用 OpenRank 评价体系计算，是一种项目内基于 Issue/PR 协作网络的计算方式，详情见 [OpenRank 介绍文档](#)。



### 大模型开发生态整体贡献度 Top 10 国家分布

#	1	2	3	4	5	6	7	8	9	10
国家	美国	中国	德国	印度	英国	加拿大	法国	波兰	荷兰	挪威
贡献比例	37.41%	18.72%	6.46%	4.25%	3.88%	3.53%	2.37%	2.16%	1.56%	1.35%
识别到的开发者数量	29451	22463	7612	9931	5711	4522	3961	1542	2144	585

### 不同技术领域下的贡献度 Top 3 国家分布

领域	AI Agent			AI Infra			AI Data			
	国家	美国	中国	德国	美国	中国	德国	美国	中国	德国
贡献比例	24.62%	21.5%	10.41%	43.39%	22.03%	3.95%	35.76%	10.77%	6.78%	

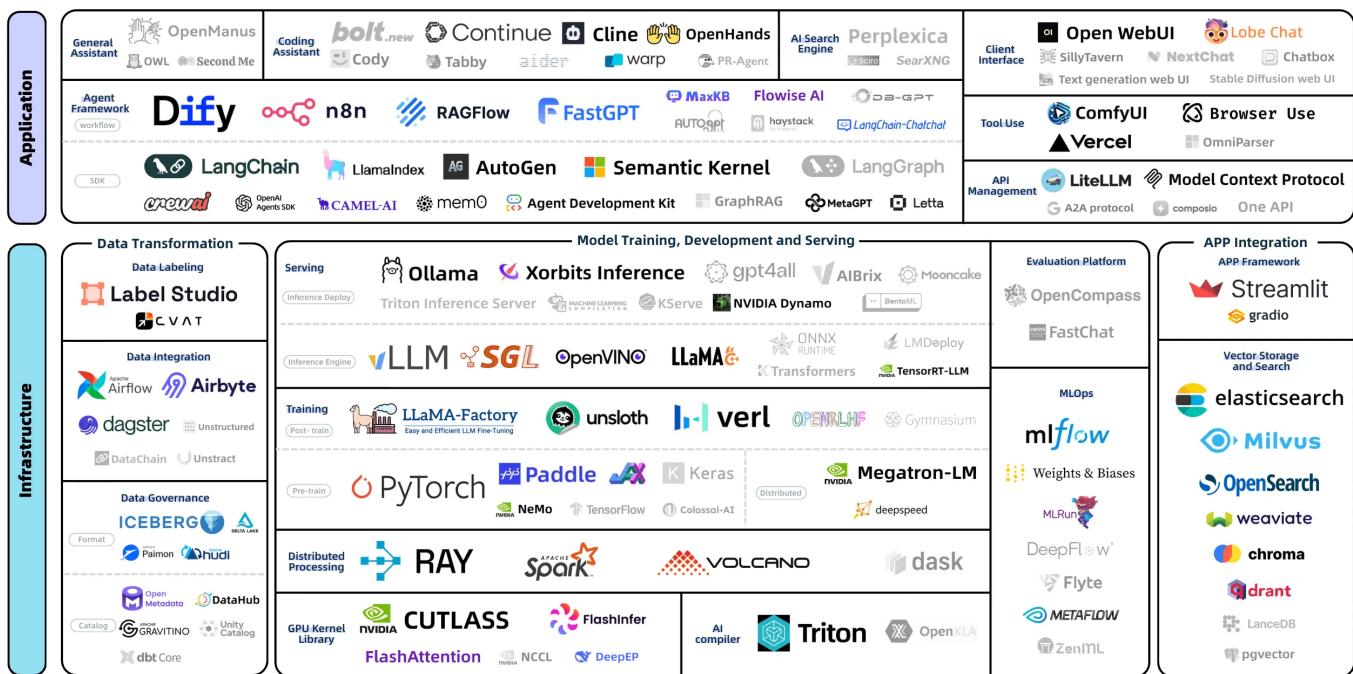
从技术领域来看，AI Infra 开发者 56,206 人，AI Agent 开发者 56,580 人，AI Data 开发者 27,018 人，这个总和与 12 万开发者的总人数相差不多，说明不同领域下的重复开发者比例不高，大多数人只参与一个技术领域下的项目。

从三大技术领域下的国家贡献度分布来看，整体以中美为主导，在 AI Infra 领域中美的领先地位更加明显，两国在基础设施领域的贡献度达到 60% 以上，排名第三的德国不足 4%，可见在基础设施领域中美有较强的控制力；AI Data 领域全球的参与情况更加平均，中美的总体贡献占比仅 46.5%，欧洲各国，如波兰、挪威、法国、荷兰等国的参与度均进入全球前十；AI Agent 领域中美差距大幅缩小，贡献度占比分别为 24.6% 和 21.5%，中国开发者在 Agent 层面相较其他领域的投入更多。

## 100 天之后的变与不变

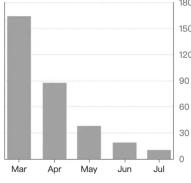
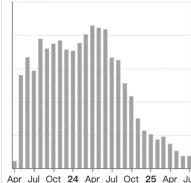
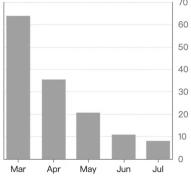
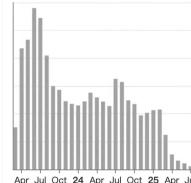
和 3 个月前的第一个版本相比，除了最明显的项目更替之外，我们也对整体的生态结构和领域做了合并、拆分和描述的调整，例如，将笼统的“Infrastructure”和“Application”的一级分类描述修改为更加具体的、也已经在逐渐发展出清晰技术边界的“AI Infra”、“AI Agent”和“AI Data”。技术仍在高速的发展，尤其在 Agent 领域，项目之间的定位和边界必然会随着技术发展而动态演化，我们可以通过 Landscape 的变化，观察到一个新的技术生态从混乱逐渐归为有序的过程。

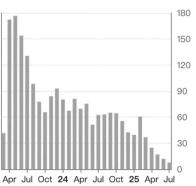
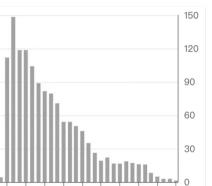
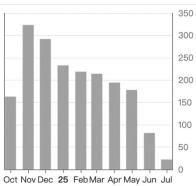
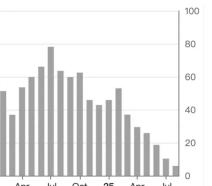
# 哪些领域和项目出局了



无论 Manus、Perplexity 这些商业产品发展和普及程度如何，在开源生态里，相关领域下的开源项目都并没有得到很好的发展。

出局的项目中，有不少可能正在步入“AI 墓园”的路上

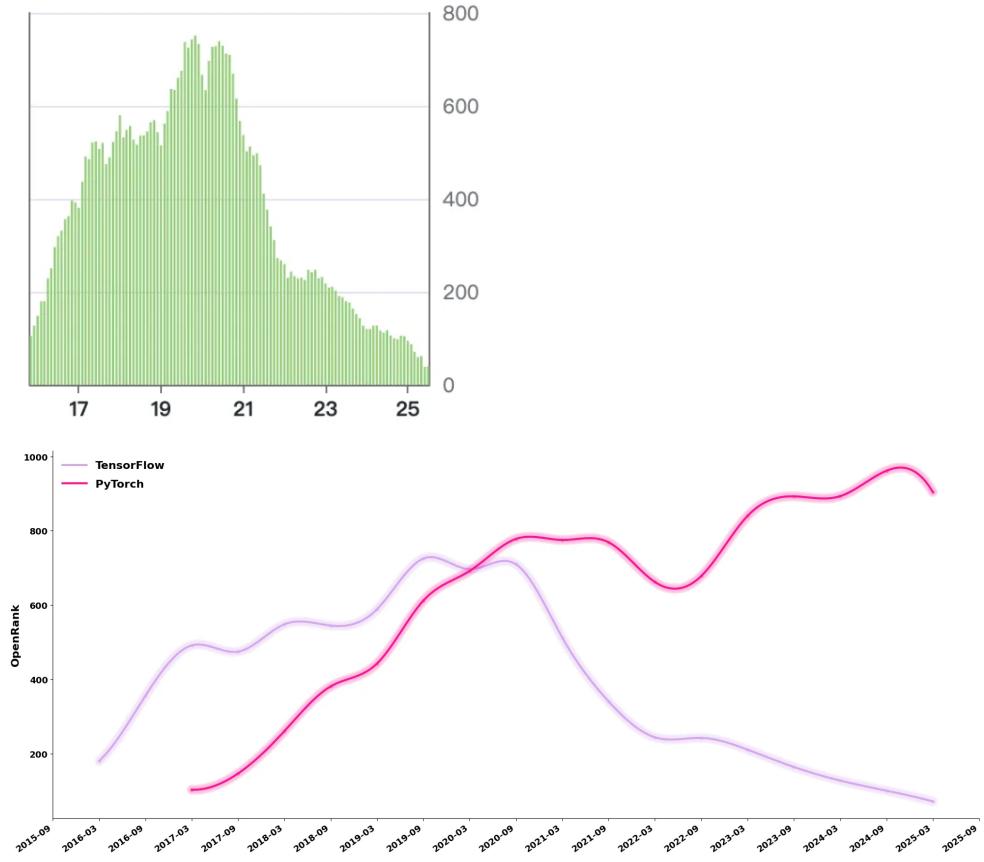
AI Agent				AI Infra			
Project	Domain	Star	Trend	Project	Domain	Star	Trend
<a href="#">OpenManus</a>	General Agent	4927		<a href="#">MLC-LLM</a>	Inference Deploy	21188	
<a href="#">OWL</a>	General Agent	17936		<a href="#">GPT4All</a>	Inference Deploy	76557	

<u>NextChat</u>	Chatbot	8566		<u>FastCh</u> at	Inference Enginee	3903	
<u>Bolt.new</u>	AI Coding	15584		<u>TGI</u>	Inference Enginee	10453	

3月份 Manus 一时爆火，多智能体框架 MetaGPT 和 Camel AI 紧随其后推出了开源版本的 OpenManus 和 OWL，但也仅仅只是昙花一现；NextChat 是最早一批流行的大模型客户端应用的项目，但后续的迭代和新特性接入速度远远比不上 Cherry Studio、LobeChat 等后起之秀，渐渐无人维护；Bolt.new 作为流行的全栈 web 开发工具，以开放模板的方式被开源出来，且很少合入外部的代码，因此项目开发者也在大幅减少。

一度非常流行的两个端侧模型部署的工具：MLC-LLM 和 GPT4All，前者绑定了自家的推理引擎 MLCEngine，后者和 Ollama 同样使用了端侧推理引擎 llama.cpp，然而最终这个生态位还是被 Ollama 拔得头筹；FastChat 是 LMSYS 在模型训练、推理和评测等环节的早期尝试，如今他们已经有了更成功的 SGLang 和 LMarena 平台，而更早出现的 TGI，由于性能落后于 vLLM 和 SGLang 等引擎，也渐渐被 HuggingFace 放弃。

## 昔日巨星 TensorFlow 的十年消亡之路



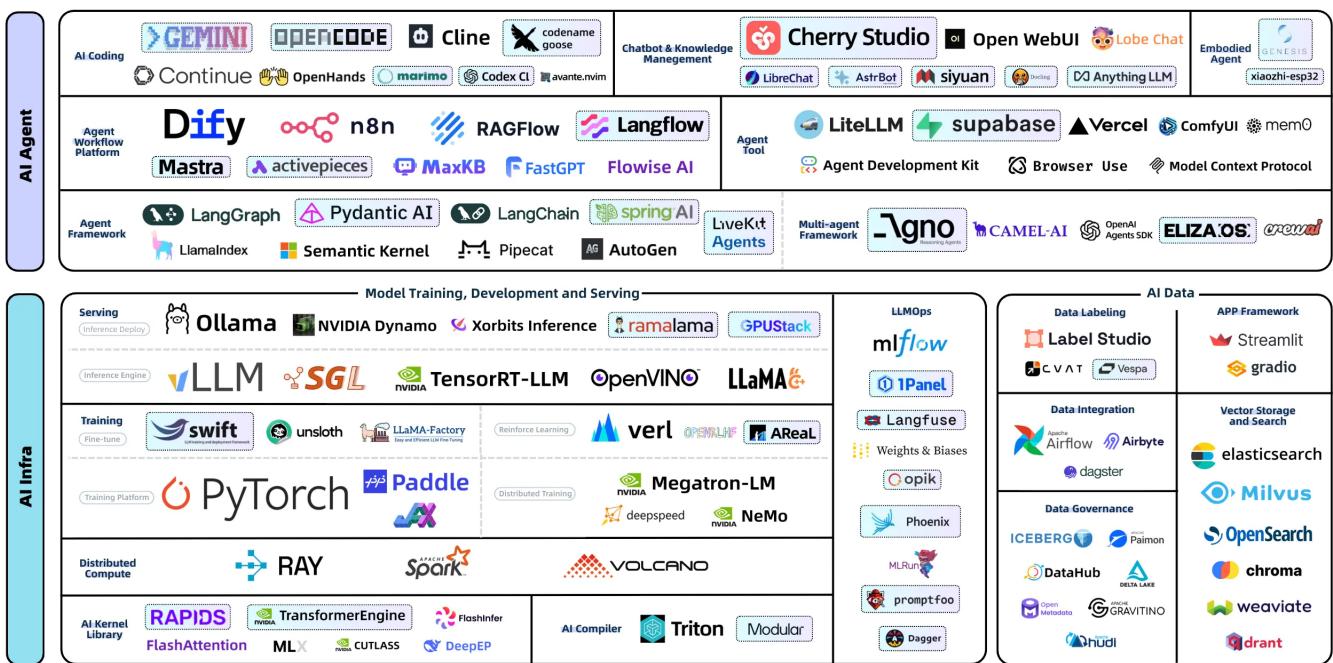
2025 年 11 月，谷歌将 TensorFlow 以 Apache 2.0 开源，很快发展为深度学习领域的主导框架。从诞生之初，TensorFlow 就为生产环境而设计，这与后来发布的 PyTorch 采取的“Pythontic”和“研究员优先”构建理念截然不同。作为开发下一代模型的创新者，研究人员倾向于选择 PyTorch，因为它灵活、易用。

2019 年 10 月，TensorFlow 发布了 2.0 版本，借鉴了 PyTorch 的核心理念，简化了模型构建。然而，这种技术上的合理转变却付出了巨大的代价：由于缺乏无缝的向后兼容性，以及复杂的迁移工具，许多已经转向 PyTorch 的开发者不愿意承担迁移遗留的 1.x 代码和学习新 API 的负担，从而对 PyTorch 的忠诚度更加坚定。正是在这个时间点，PyTorch 社区正式超过了 TensorFlow，两个项目也从此走向了分化的发展曲线。

其他代表性的被拿掉的，下回再书

- pgvector/pgvector 始终不温不火的向量中间件
- microsoft/graphrag 社区数据节节败退，微软回撤在 GraphRAG 的布局

## 哪些领域和项目第一次进入视野



领域的变化主要体现在 Agent 层面，以 AI Coding、Chatbot 和开发框架为主的领域出现了不少新的高热度项目。在其中，还发现了两个和具身智能应用场景相关有趣项目：

- 小智 AI 聊天机器人：构建一个基于 ESP32 微控制器的 AI 语音交互设备——“AI 小智”，让大语言模型（如 Qwen、DeepSeek）能运行在硬件中。
- Genesis：面向通用机器人与具身的物理仿真平台，用途包括机器人学习、物理模拟、渲染与数据生成，具备极高的科研与应用价值。

Infra 层面在领域的变化主要体现在对“模型运维”这一概念的整合，我们将原先涉及到模型评测和传统机器学习运维的领域合并在一起，成为纵穿模型全生命周期的 LLMOps，它本质上是 MLOps 在大语言模型时代的延伸，解决的是如何在真实生产环境下高效、可靠、可控地使用 LLM。当前 LLMOps 领域下的这些项目覆盖了模型与应用的可观测性（langfuse、phoenix）、模型评测与基准测试（promptfoo）、agentic workflow 的运行时环境管理（1Panel、dagger）等环节。

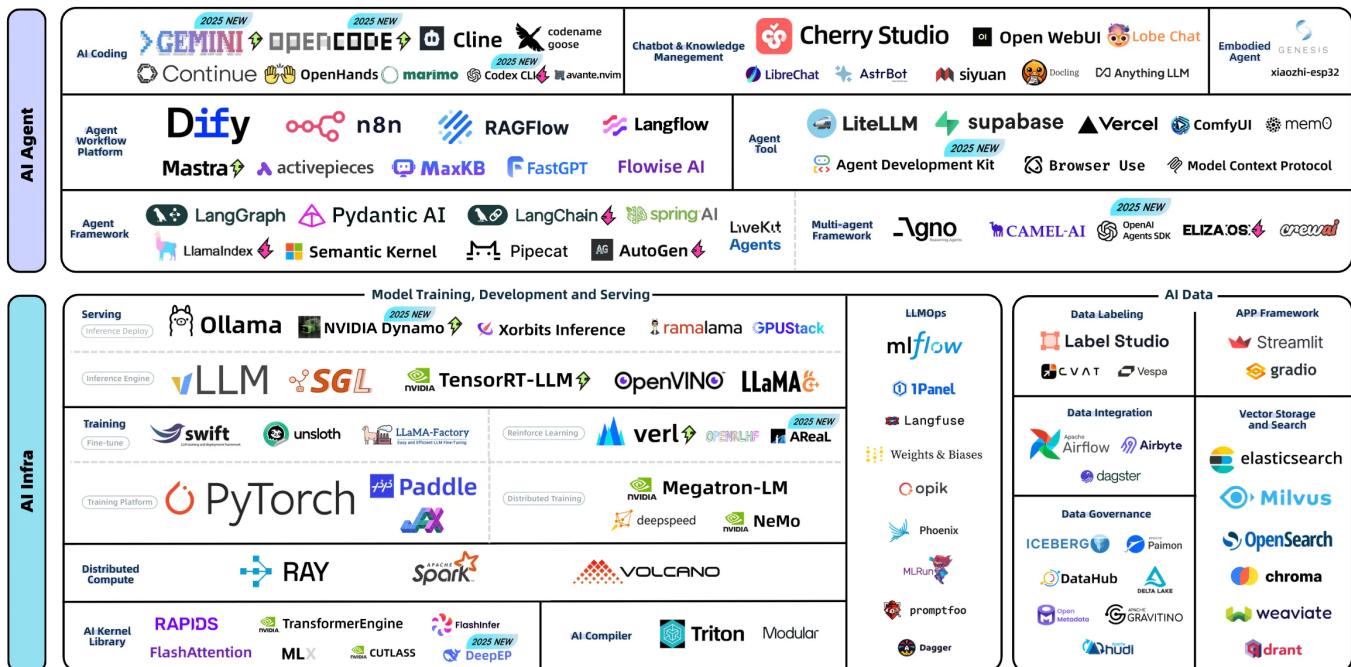
## 新近项目中的 OpenRank Top 10

#	Project	Domain	OpenRank	Stars	Trend	Language	Creation	Initiator	License
1	google-gemini/genie-mini-cli	AI Coding	391	668 81	↗	TypeScript	2025-04-17	Google	Apache 2.0

2	 <a href="#">CherryHQ/cherry-studio</a>	Chatbot	313	309		TypeScript	2024-05-24	Cherry Studio	User-Segment Dual Licensing
3	<a href="#">sst/opencode</a>	AI Coding	195	1707		Go	2025-04-30	sst.dev	MIT
4	<a href="#">supabase/supabase</a>	Agent Tool	178	865		TypeScript	2019-10-12	Supabase	Apache 2.0
5	<a href="#">langflow-ai/langflow</a>	Agent Platform	143	9512		Python	2023-02-08	IBM DataStage (Acquired)	MIT
6	<a href="#">block/goose</a>	AI Coding	139	1812		Rust	2024-08-23	Block Inc	Apache 2.0
7	<a href="#">mastra-ai/mastra</a>	Agent Platform	135	1551		TypeScript	2024-08-06	Mastra	Apache 2.0
8	<a href="#">modelscope/ms-swift</a>	Fine-tune	113	906		Python	2023-08-01	AlibabaCloud	Apache 2.0
9	<a href="#">agnosagi/agno</a>	Multi-agent Framework	106	3121		Python	2022-05-04	Agno	MPL 2.0
10	<a href="#">modular/modular</a>	AI Compiler	98	245		Mojo	2023-04-28	Modular	Apache 2.0

其中，终端 AI 编程助手 Gemini CLI 和模型客户端交互聊天工具 Cherry Studio 还在本次大模型全景图的所有项目中位列第 3 和第 7。

# 没变的是：此消彼长，前浪后浪，增长与衰落，一如既往



## 全景图上的「the new wave」

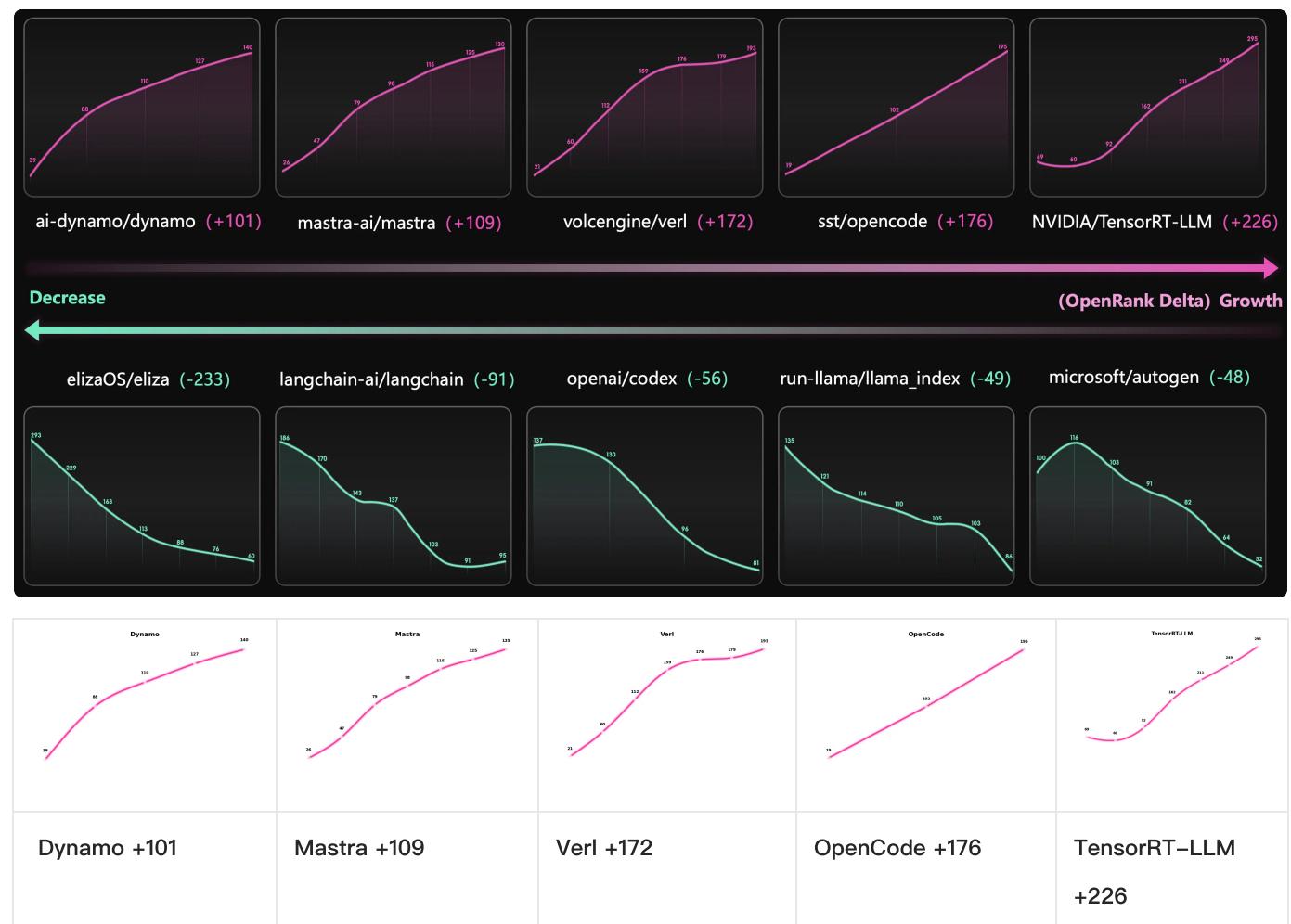
Project	Domain	OpenRank	Stars	Language	Creation	Initiator	License
<a href="#">OpenCode</a>	AI Coding	195	17071	Go	2025-04-30	Anomaly Innovations Inc	MIT
<a href="#">Gemini CLI</a>	AI Coding	391	66881	TypeScript	2025-04-17 (5月仓库公开)	Google	Apache 2.0
<a href="#">Codex CLI</a>	AI Coding	81	35282	Rust	2025-04-13	OpenAI	Apache 2.0
<a href="#">adk-python</a>	Agent Tool	109	11477	Python	2025-04-01	Google	Apache 2.0
<a href="#">openai-agents-python</a>	Multi-agent Framework	72	13225	Python	2025-03-11	OpenAI	MIT
<a href="#">Dynamo</a>	Inference Deploy	140	4642	Rust	2025-03-03	NVIDIA	Apache 2.0

在这些 2025 发起的新势力项目中，OpenCode 来自于创业公司 Anomaly Innovations，并且在发起之日就定位为是 Claude Code 的 100% 开源替代。在剩下的几个项目中，我们可以看出 OpenAI、Google、NVIDIA 这些大厂通过开源开放的工具链来建立围绕其闭源模型或硬件生态的护城河的野心：Dynamo 在支持 vLLM、SGLang 和自家的 TensorRT-LLM 等主流推理后端的同时，也完美适配 NVIDIA GPU 的硬件特性，在成为高吞吐、多模型部署的行业级工具之后，会进一步促使企业倾向选择 NVIDIA 硬件以最大化性能收益。

OpenAI 和 Google 通过在应用层布局开源工具链，把开发者逐渐锁定在各自闭源模型的生态之中的野心则更加明显：adk-python 和 openai-agents-python 是专为其各自的闭源模型封装的 Agent 系统构建工具，Google 甚至做了云服务的生态优化，支持在 Google Cloud 上优先部署编排好的智能体；而 Codex CLI 和 Gemini CLI 同样效仿了 Claude Code 这种在终端实现高度自治的代码理解与修改的形态，把大模型直接带到开发者最熟悉的命令行里，一个深度绑定 Gemini，一个兼容 OpenAI 并开放 MCP 接口。

在接下来的一段时间，我们可以拭目以待，看看这些项目是否达到了它们被寄予的战略使命。

### 全景图上的「up and down」



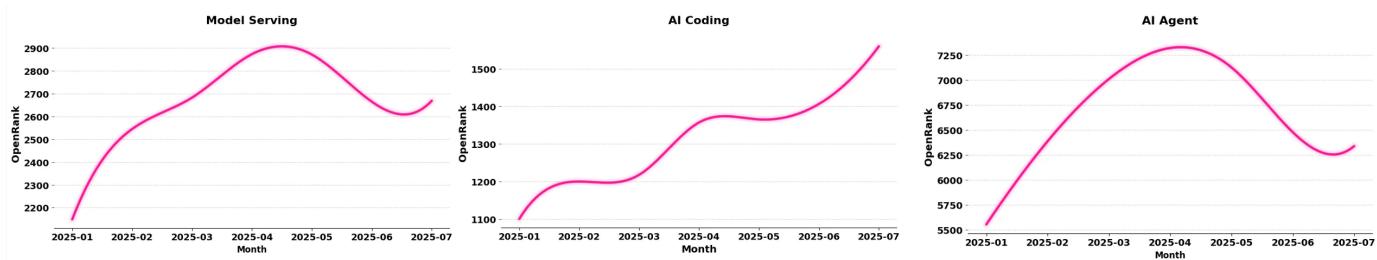


上述十个项目，分别在近半年内 OpenRank 的增长和下降绝对值与比例都位列前茅的项目，图上我们展示的是他们从 2 月到 8 月的 OpenRank 绝对值变化。

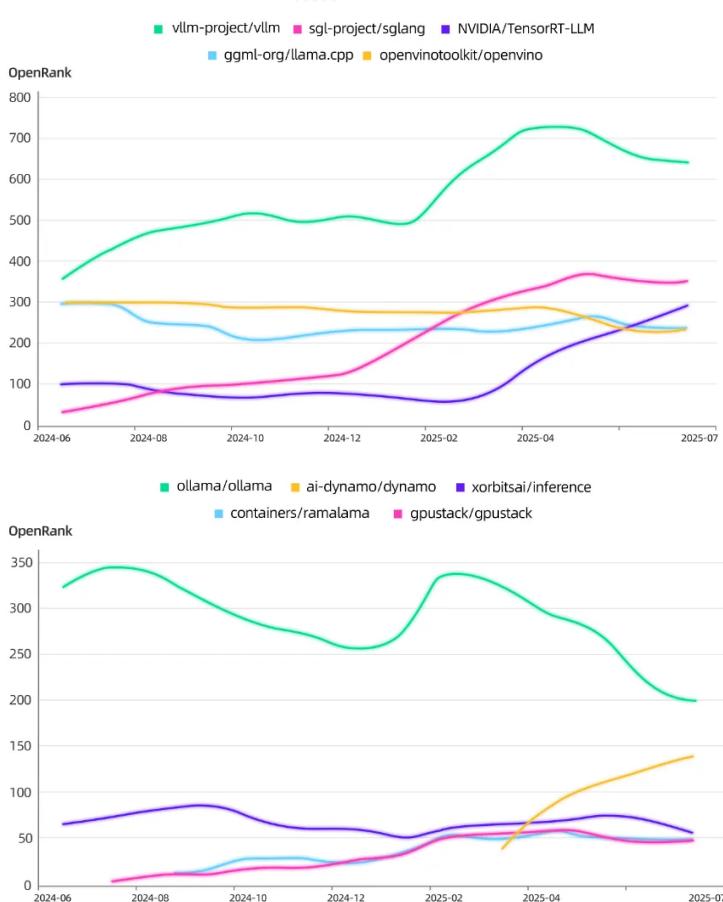
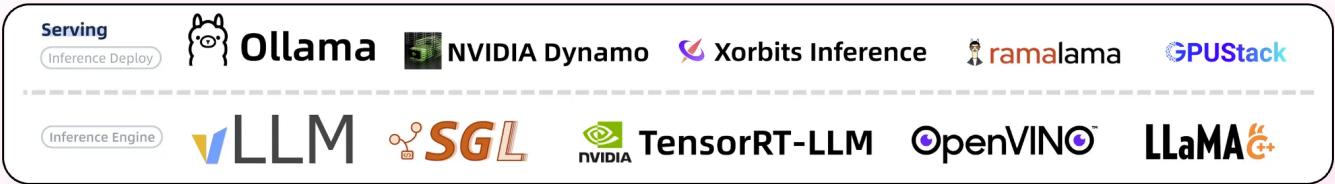
增长较明显的五个项目，分别是：NVIDIA 推出的企业级推理引擎后端 [TensorRT-LLM](#) 和多租户推理编排工具 [Dynamo](#)、字节推出 LLM 强化学习框架 [verl](#)、对标 Claude Code 的开源命令行 Coding 工具 [OpenCode](#)、面向 TypeScript/JavaScript 开发者的 Agent 编排框架 [Mastra](#)。

下降较明显的五个项目，有四个都是 Agent 编排框架：[Eliza](#)、[LangChain](#)、[LlamaIndex](#) 和 [AutoGen](#)。剩下的一个项目，是 OpenAI 在 4 月新推出的命令行 Coding 工具 [Codex](#)，相较于 Gemini CLI 的快速增长，它的起步看起来有一些出师不利。

## 专题洞察：技术趋势下的项目故事



### 第一篇：Model Serving



模型服务的本质，是把训练完成的大模型以一种可被应用层稳定调用的方式运行起来。它需要解决的不仅仅是“能不能跑”，更是“能不能高效、可控地跑”。在场景上，**大规模在线推理**是模型服务的主战场，数据中心级的部署支撑了数以千万计的请求；与此同时，企业内部也常常出于安全与合规的考量，搭建私有化的推理服务；而在端侧，像 llama.cpp 或 Ollama 这样的项目，让模型能在个人电脑、手机甚至嵌入式设备上运行，满足离线和隐私需求；还有越来越多的混合模式，部分处理在端侧完成，复杂推理则交给集群完成。

2023 年以来的快速演进已经让模型服务成为连接 AI 基础设施与应用层的关键中间件。一方面，vLLM、SGLang 等代表性的推理引擎项目不断在高吞吐、长上下文、多用户并发的场景里打磨出极致性能，另一方面，Ollama、llama.cpp 等则推动了本地可用性和生态扩散，让大模型“跑在你手边”成为现实。同时，NVIDIA Dynamo 这样的编排框架正在把单机高效推理扩展到多节点、多模型、多租户的集群层面。

### 集群部署：NVIDIA Dynamo

- 技术定位：**定位企业级推理编排层，解耦 prefill/decode 并优化 GPU 路由，以多模型、多后端、多节点集群调度为核心。
- 开源社区：**2025 年开源，由 NVIDIA 主导，仍处于早期阶段，但因其与 TensorRT-LLM、vLLM 等深度绑定，迅速获得产业级关注。

### 端侧部署：Ollama

- 技术定位：**聚焦本地与端侧部署，基于 llama.cpp 内核提供一键运行与 OpenAI 兼容 API，让大模型“先跑起来”更简单。
- 开源社区：**自 2023 年以来快速走红，凭借易用性吸引开发者与应用生态广泛接入，形成活跃的本地化开发社区。

### 集群推理：vLLM SGLang

- 技术定位：**面向高吞吐集群推理，极致压榨 GPU 性能，均支持生态中的主流算子库，并融合了并行、量化等关键能力。
- 开源社区：**均出自学术/开源先锋团队 (UC Berkeley、LMSys)，2023–2024 年星标与贡献者增长显著，已成为科研与产业部署的事实标准。

### 端侧推理：LLaMA

- 技术定位：**以极致轻量和多平台移植为目标，依靠多种量化与硬件后端优化，实现 CPU、GPU、移动端乃至浏览器的本地推理。
- 开源社区：**2023 年起成为端侧推理的代名词，依靠极低依赖与广泛适配，衍生出大量工具与下游项目，形成了极具韧性的社区生态。

### 集群部署：Dynamo

- 技术定位：定位企业级推理编排层，解耦 prefill/decode 并优化 GPU 路由，以多模型、多后端、多节点集群调度为核心。
- 开源社区：2025 年开源，由 NVIDIA 主导，仍处于早期阶段，但因其与 TensorRT-LLM、vLLM 等深度绑定，迅速获得产业级关注。

### 端侧部署：Ollama

- 技术定位：聚焦本地与端侧部署，基于 llama.cpp 内核提供一键运行与 OpenAI 兼容 API，让大模型“先跑起来”更简单。
- 开源社区：自 2023 年以来快速走红，凭借易用性吸引开发者与应用生态广泛接入，形成活跃的本地化开发社区。

### 集群推理：vLLM, SGLang

- 技术定位：面向高吞吐集群推理，极致压榨 GPU 性能，均支持生态中的主流算子库，并融合了并行、量化等关键能力。
- 开源社区：均出自学术/开源先锋团队 (UC Berkeley、LMSys)，2023–2024 年星标与贡献者增长显著，已成为科研与产业部署的事实标准。

### 端侧推理：llama.cpp

- 技术定位：以极致轻量和多平台移植为目标，依靠多种量化与硬件后端优化，实现 CPU、GPU、移动端乃至浏览器的本地推理。
- 开源社区：2023 年起成为端侧推理的代名词，依靠极低依赖与广泛适配，衍生出大量工具与下游项目，形成了极具韧性的社区生态。

## 第二篇：AI Coding



从最初的单一代码补全功能发展到如今的多模态支持、上下文感知与协同工作流，AI Coding 的核心技术在不断进化。CLI 工具如 Gemini CLI 和 OpenCode 利用 AI 模型的强大推理能力，将开发者的需求转化为更高效的编程体验；与此同时，插件形态的工具，如 Cline 和 Continue，通过无缝集成到现有开发平台中，让开发者在保持现有工作流的基础上享受 AI 提供的各种智能服务，极大地提升了开发效率。Goose 和 OpenHands 等协作开发平台，将 AI 能力融入团队项目管理、代码审查、任务分配等各个环节，推动了跨地域、跨职能的团队协作。而 Claude Code、Cursor 和 Windsurf 等闭源的商业化项目，也吸引了大量个人开发者和企业客户。随着市场需求的提升，AI Coding 的商业化潜力巨大，付费订阅、SaaS 服务、增值功能等将成为未来的主要盈利模式。

### CLI 形态

**Open Source:**

**Closed Source:**

大厂下场，通过和自家模型深度绑定，把开发者逐渐锁定在各自闭源模型的生态之中。今年 Claude Code 推出之后，这种形态开始成为开发者的最新优选。

### IDE 形态

**Open Source:**

**Closed Source:**

商业化售卖为主，Cursor、Windsurf 验证了市场热情，国内大厂字节、阿里也纷纷下场。Marimo 定位为 AI 原生的交互式 notebook 编辑器，是看到的为数不多的开源 IDE。

### 插件形态

创业团队为主，通过无缝集成到现有的 IDE 和编辑器等开发环境中，让开发者在保持现有工作流的基础上，享受 AI 提供的智能服务。

### 协作开发工作流

创业团队为主，将 AI 能力融入项目管理、协作开发、代码审查等环节，服务场景从单纯的辅助个人开发者编写代码扩展到企业级研发效能管理的开发环境中。

### CLI 形态

- **开源：**Gemini CLI, OpenCode, Codex CLI
- **闭源：**Claude Code

大厂下场，通过和自家模型深度绑定，把开发者逐渐锁定在各自闭源模型的生态之中。今年 Claude Code 推出之后，这种形态开始成为开发者的最新优选。

### IDE 形态

- **开源：**Marimo
- **闭源：**Cursor, Windsurf, Trae, Qoder

商业化售卖为主，Cursor、Windsurf 验证了市场热情，国内大厂字节、阿里也纷纷下场。Marimo 定位为 AI 原生的交互式 notebook 编辑器，是看到的为数不多的开源 IDE。

## 插件形态

### Cline, Continue, Avante.nvim

创业团队为主，通过无缝集成到现有的 IDE 和编辑器等开发环境中，让开发者在保持现有工作流的基础上，享受 AI 提供的智能服务。

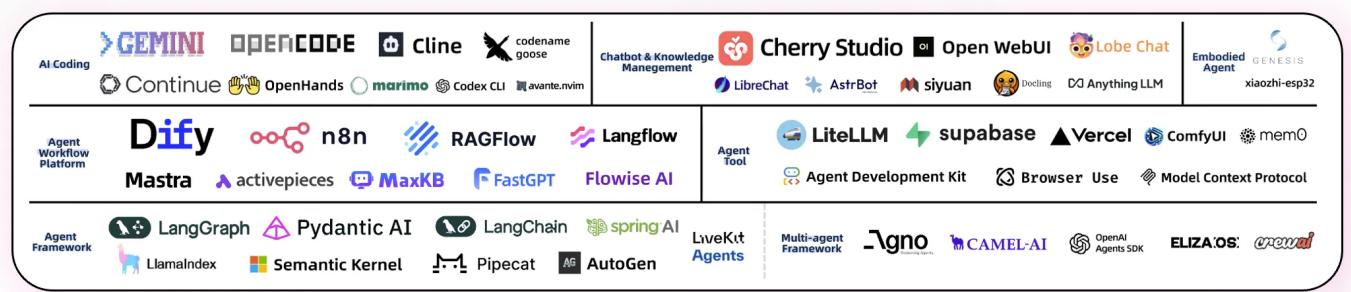
## 协作开发工作流

### Goose, OpenHands

创业团队为主，将 AI 能力融入项目管理、协作开发、代码审查等环节，服务场景从单纯的辅助个人开发者编写代码扩展到企业级研发效能管理的开发环境中。

## 第三篇：AI Agent

人们常说 2025 年会是 AI 应用真正落地的一年。最初，是 LangChain、LlamaIndex 等框架提供了基础的 Agent 搭建方式；随后，开源生态中出现了专注于不同环节的项目，如 Mem0（记忆）、Browser-Use（工具调用）、Dify（工作流执行）、LobeChat（交互界面），开源社区正在构建完整的拼图，为更强大的自治 AI 系统打下基础。每个项目聚焦的方向不同，但目标一致：让 AI 更加智能地理解、记忆、行动和交互，从而真正解放生产力。



<p><b>记忆:</b>  mem0</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 作为 AI 记忆系统，提供跨会话、跨任务的智能记忆管理，通过向量数据库和图数据库等多种技术实现对用户与 Agent 的个性化记忆，提升对话的连贯性和上下文处理能力。</li> <li><b>开源社区:</b> 拥有 39k Star 和 200+ 贡献者，在 AI 助手、客服机器人等领域获得广泛应用。</li> </ul>	<p><b>执行（工作流）:</b>  Dify</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> AI Agent 能够自动化操作浏览器，执行如信息抓取、表单填写等复杂的网页任务，突破了传统爬虫和脚本的局限，增强了 AI 系统对互联网的动态交互能力。</li> <li><b>开源社区:</b> 拥有 69k Star 和 200+ 贡献者，在网页自动化和 RPA 领域得到了广泛应用，并推动了对大模型与浏览器自动化结合的讨论和发展。</li> </ul>
<p><b>工具:</b>  Browser Use</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 全栈式的 LLM 应用开发平台，集成了任务编排、模型管理、知识检索等功能，帮助开发者构建、部署和管理智能 Agent，面向企业级应用提供了高效的工具链和全流程支持。</li> <li><b>开源社区:</b> 拥有 113k Star 和 900+ 贡献者。支持与多种主流 LLM、插件、API 的集成，成为 LLMOps 和 Agent 开发的重要生态平台，吸引了大量企业合作和生态扩展。</li> </ul>	<p><b>交互:</b>  Lobe Chat</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 支持多智能体和多模态交互的开源框架，允许用户创建、切换和组合多个 Agent，以适应不同任务需求，专注于用户友好的对话体验和协作式工作流。</li> <li><b>开源社区:</b> 拥有 65k Star 和 250+ 贡献者，凭借其易用的 UI 和强大的插件机制，吸引了开发者构建自己的 Agent 模块，并在智能助手、创作工具等领域获得了广泛的社区支持和活跃贡献。</li> </ul>

## 规划: LLM + Prompt

<p><b>记忆:</b> mem0</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 作为 AI 记忆系统，提供跨会话、跨任务的智能记忆管理，通过向量数据库和图数据库等多种技术实现对用户与 Agent 的个性化记忆，提升对话的连贯性和上下文处理能力。</li> <li><b>开源社区:</b> 拥有 39k Star 和 200+ 贡献者，在 AI 助手、客服机器人等领域获得广泛应用。</li> </ul>	<p><b>工具:</b> Broswer use</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 使 AI Agent 能够自动化操作浏览器，执行如信息抓取、表单填写等复杂的网页任务，突破了传统爬虫和脚本的局限，增强了 AI 系统对互联网的动态交互能力。</li> <li><b>开源社区:</b> 拥有 69k Star 和 200+ 贡献者，在网页自动化和 RPA 领域得到了广泛应用，并推动了对大模型与浏览器自动化结合的讨论和发展。</li> </ul>
<p><b>执行（工作流）:</b> Dify</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 全栈式的 LLM 应用开发平台，集成了任务编排、模型管理、知识检索等功能，帮助开发者构建、部署和管理智能 Agent，面向企业级应用提供了高效的工具链和全流程支持。</li> <li><b>开源社区:</b> 拥有 113k Star 和 900+ 贡献者。支持与多种主流 LLM、插件、API 的集成，成为 LLMOps 和 Agent 开发的重要生态平台，吸引了大量企业合作和生态扩展。</li> </ul>	<p><b>交互:</b> Lobe-chat</p> <ul style="list-style-type: none"> <li><b>技术定位:</b> 支持多智能体和多模态交互的开源框架，允许用户创建、切换和组合多个 Agent，以适应不同任务需求，专注于用户友好的对话体验和协作式工作流。</li> <li><b>开源社区:</b> 拥有 65k Star 和 250+ 贡献者，凭借其易用的 UI 和强大的插件机制，吸引了开发者构建自己的 Agent 模块，并在智能助手、创作工具等领域获得了广泛的社区支持和活跃贡献。</li> </ul>

## 规划: LLM+prompt

## 扩展阅读

- 全景图地址: <https://antoss-landscape.my.canva.site/>
- 全景图涉及到的数据集和所有洞察内容已经开源, 欢迎提交反馈:
  - <https://github.com/antgroup/l1m-oss-landscape>
- 消费 OpenRank 及开源生态指标数据, 欢迎了解 OpenDigger:
  - <https://github.com/X-lab2017/open-digger>

小雅 394185