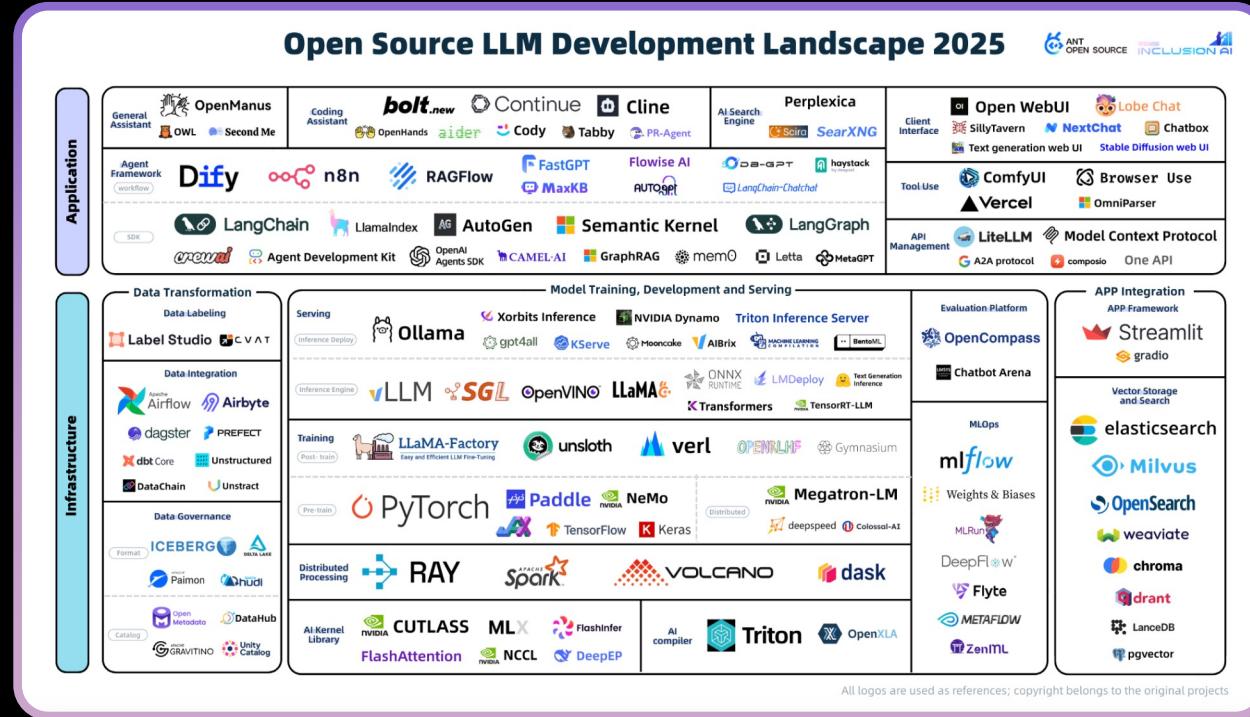


从社区数据看全球大模型开源开发生态的全景与趋势

演讲人：王旭，蚂蚁开源技术委员会副主席
时间：2025-09-13



100 天前，我们发布了一份大模型开发生态的全景图



这是一份利用开源社区中的数据制作的全景图，在寻找这个生态领域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。

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。也包含一些灵魂发问：



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？

蚂蚁做这件事情的初衷，并不是为了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，通过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，所以我们也开启了 GitHub 上一条 Issue 作为反馈的公共入口：<https://github.com/antgroup/lm-oss-landscape>

在大家反馈的基础上，我们对看生态全景的方法进行了更新



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。

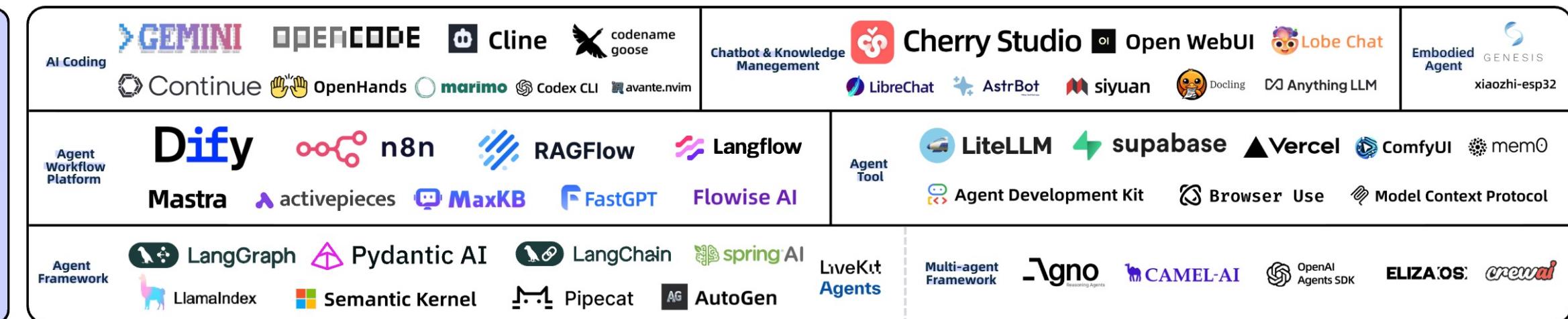
注：OpenRank 是一种社区导向的算法，在直接面向开发者、或者开发者群体基数本身就比较大的领域，值往往是更高的。但对于更底层、开发者基数更小的项目，例如算子库、编译器等，则相对难以体现出项目的价值，因此，我们不可避免地仍旧引入了一些人为判断的成分。

Open Source LLM Development Landscape

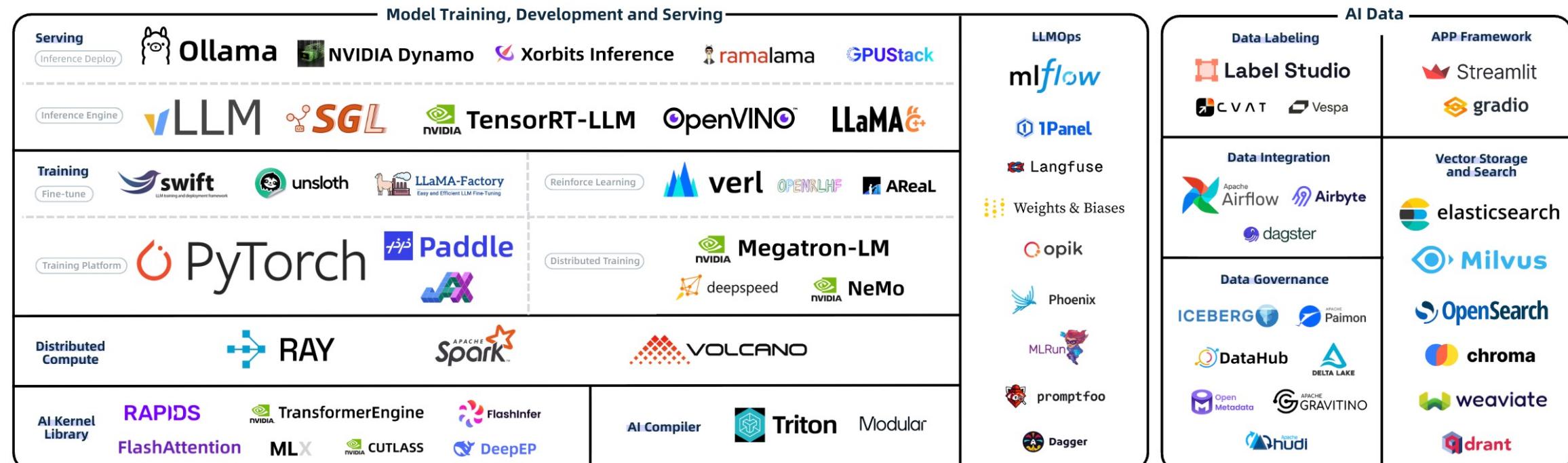
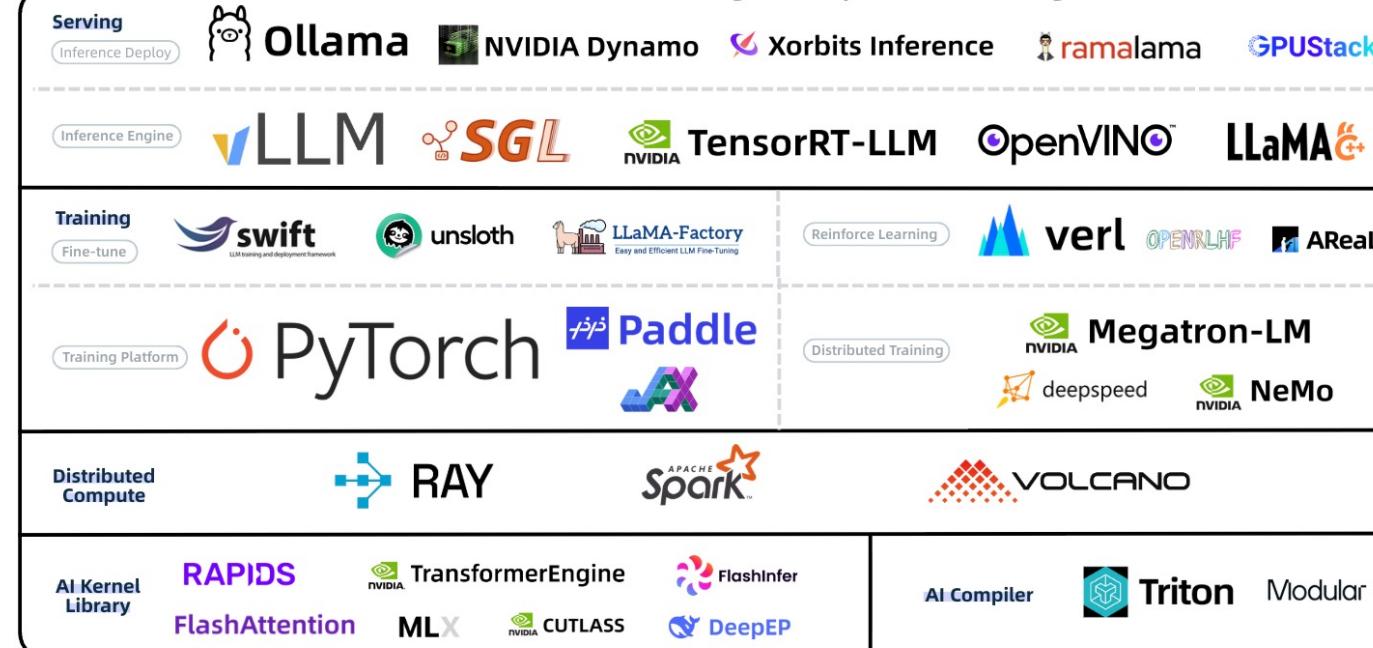


AI Agent

AI Infra



Model Training, Development and Serving



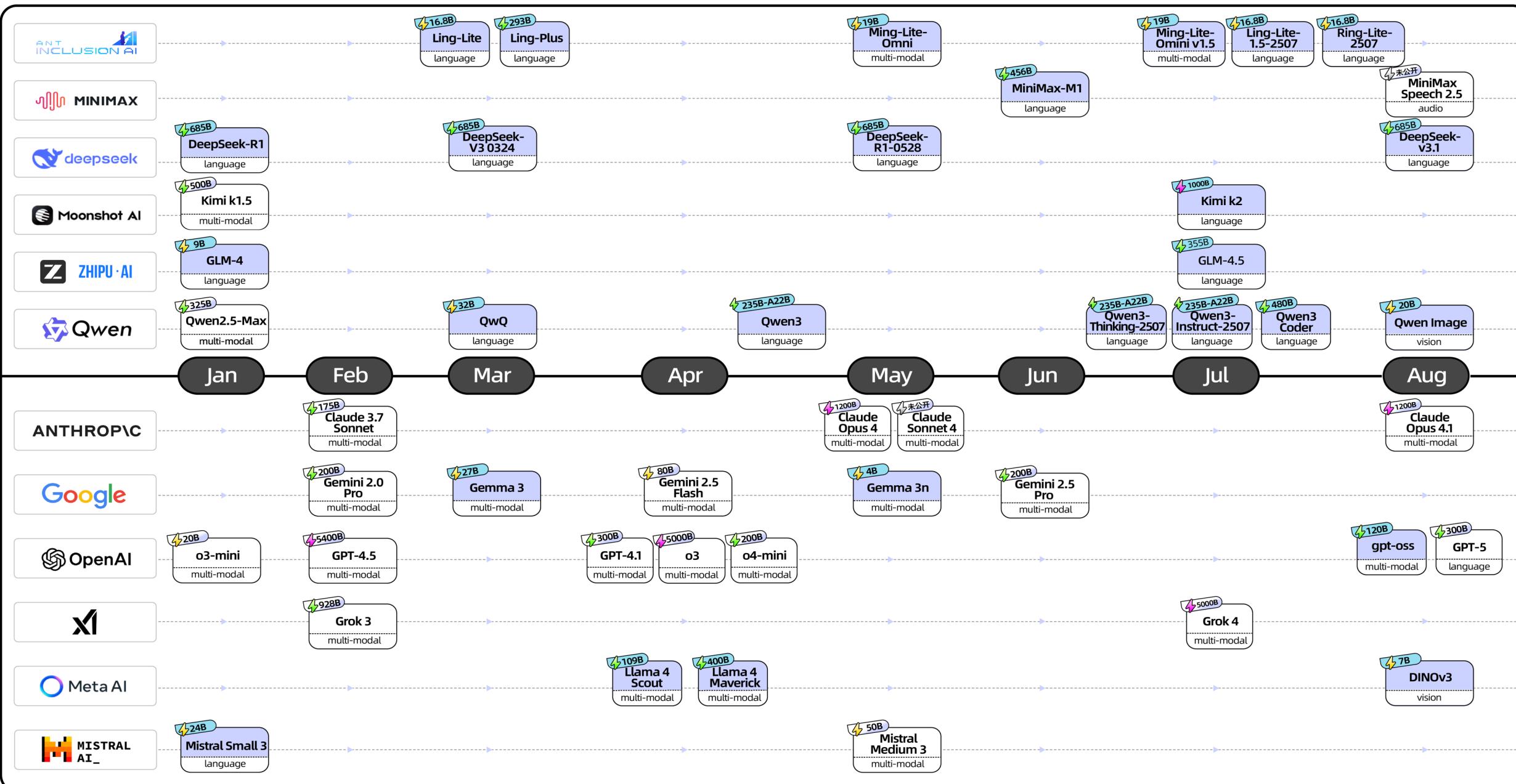
大模型开发生态总览

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** 这个新的评价标准了，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。

算上那些被拿掉的在大模型开发生态的项目们，这些项目从创建至今的“年龄”中位数是 30 个月，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%）。无论是模型的研发还是围绕着模型的开源开发生态，中美两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。

Large Models Landscape 2025



中美开源与闭源的路线分化



- 从图上能够更直观的感受中国开源大模型的百花齐放，而国外顶尖的模型厂商依旧走的是闭源路线；
- 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”元素），也正在考虑向闭源转型。

MoE 架构下模型参数在规模化发展



- DeepSeek、Qwen、Kimi 等旗舰模型全面采用了 MoE 架构，在这种架构下，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，模型是否具备 Reasoning 的能力也成为了时尚单品；
- 推理模型普遍需要更久的时间和更多的 token 来返回答案，Qwen、Claude、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



- 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，开发生态中也出现了围绕着语音模态的丰富工具链，如 Pipecat、LiveKit Agents 等；
- 在 2024 年年初，OpenAI 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。

主观和客观的不同模型评价方式



- 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，代表评测集：AIME、GPQA、SWE-bench、LiveCode Bench



Design Arena

Join 116,158 voters to discover which AI is the best at design.



LMArena




Learn how your votes power transparent AI progress

图片仅为引用，来源：<https://www.designarena.ai/>

图片仅为引用，来源：<https://lmarena.ai/>

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	Coding	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
MMMU	Multimodal	GPT-5, Claude Opus 4.1

注：上表梳理的是最近两个月新发布的模型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

再看生态与趋势



全景图上最活跃的开源项目 Top 10



#	Project	Delta	Domain	OpenRank	Star	OpenRank Trend	Language	Created	Initiator	License
1	PyTorch	-	Training Platform	859	92,039		Python	2016-08-13	Meta	BSD-3-Clause
2	LLM	-	Inference Engine	637	53,912		Python	2023-02-09	Berkeley	Apache 2.0
3	GEMINI	new	AI Coding	391	66,881		TypeScript	2025-04-17	Google	Apache 2.0
4	Dify	-1	Agent Platform	379	109,674		TypeScript	2023-04-12	Dify.AI	Open Source License
5	SGL	+3	Inference Engine	352	16,555		Python	2024-01-08	LMSYS	Apache 2.0
6	Apache Airflow	+1	Data Integration	323	41,371		Python	2015-04-13	Airbnb	Apache 2.0
7	Cherry Studio	new	Chatbot	313	30,976		TypeScript	2024-05-24	Shanghai Qianhui Technology	User-Segmented Dual License
8	TensorRT-LLM	new	Inference Engine	295	11,219		C++	2023-08-16	NVIDIA	Apache 2.0
9	n8n	+6	Agent Platform	281	126,659		TypeScript	2019-06-22	n8n	Sustainable Use License
10	RAY	+2	Distributed Compute	274	38,302		Python	2016-10-25	Anyscale	Apache 2.0

注：以上数据截止 2025 年 8 月 1 日

如何定义这个时代的开源？



多数 LLM 开发生态的项目仍然采用的是 Apache 2.0 或 MIT 宽松许可，但在 Top 10 的项目中，仍有一些值得关注的特殊案例：



Dify 的 Open Source License

基于 Apache 2.0 许可的文本做了修改，增加了两个附加条款：1. 限制未经许可的多租户环境运营；2. 使用 Dify 前端时，不得移除或修改 Logo 和版权信息。



n8n 的 Sustainable Use License

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：1. 仅限于在企业内部，或者非商业、个人用途下使用或修改；2. 在分发时，必须基于非商业目的免费提供；3. 不能更改软件中的许可、版权或作者信息。

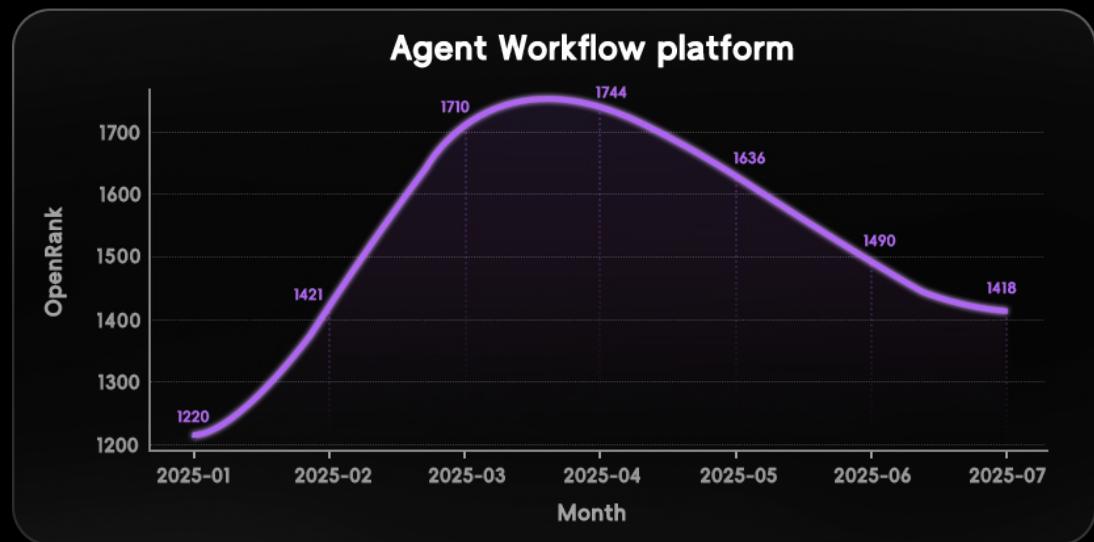
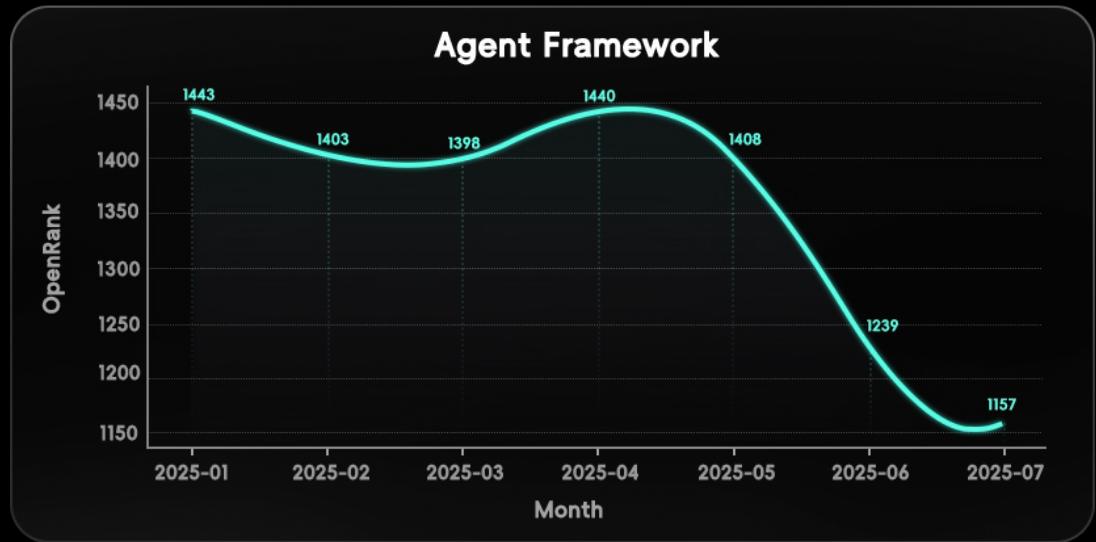
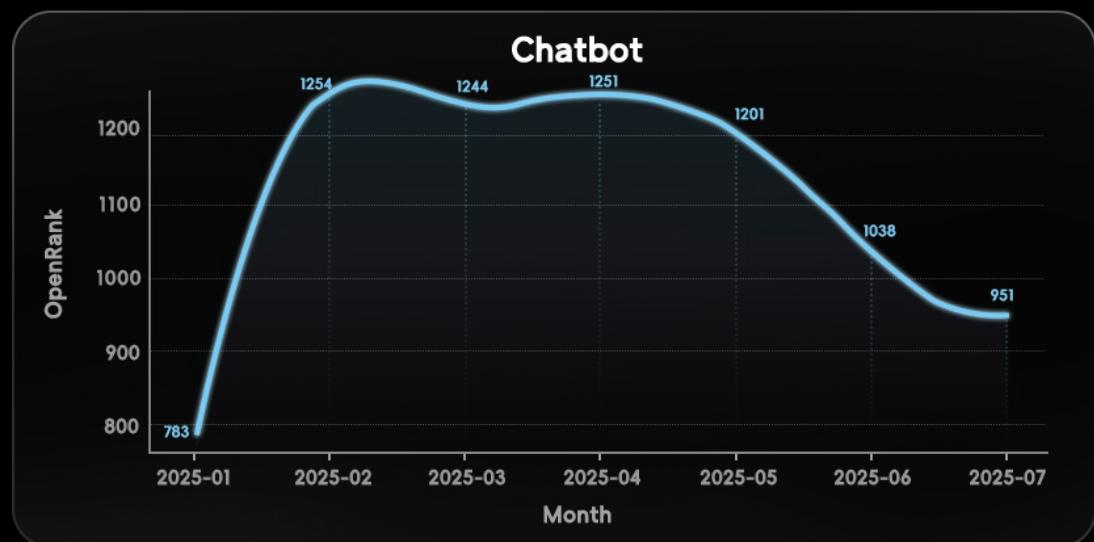
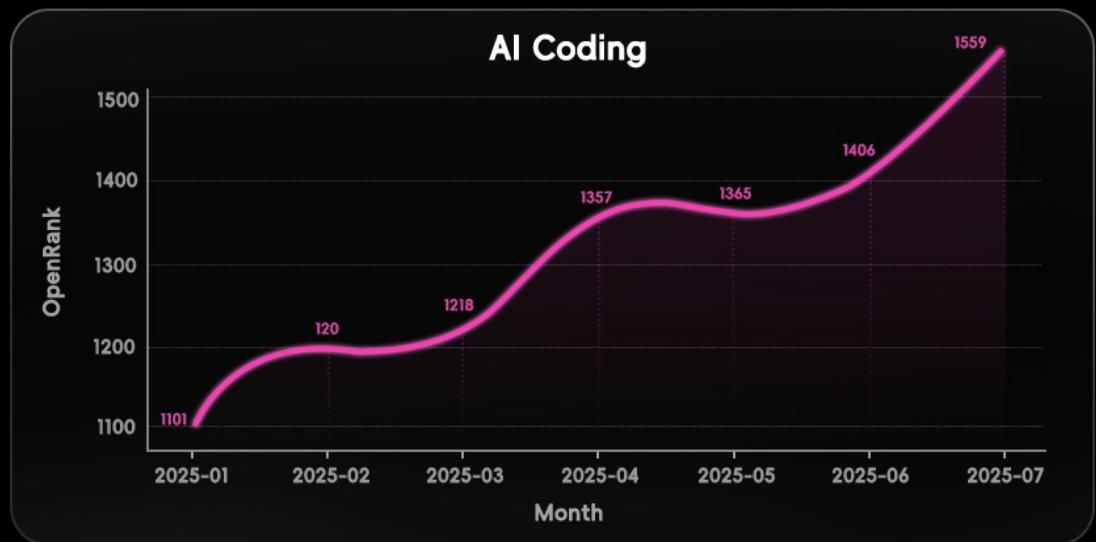


Cherry Studio 的 User-Segmented Dual License

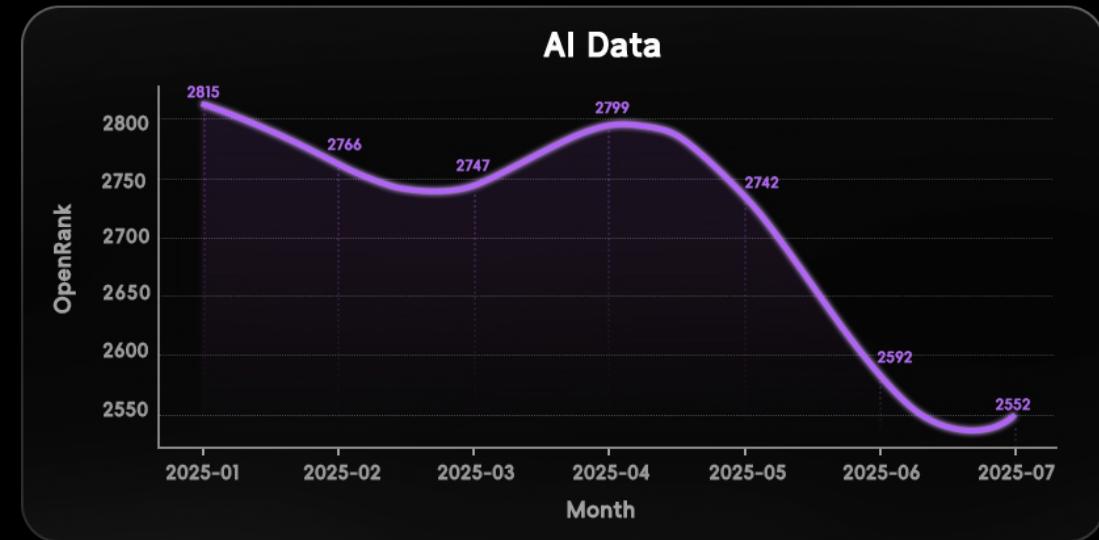
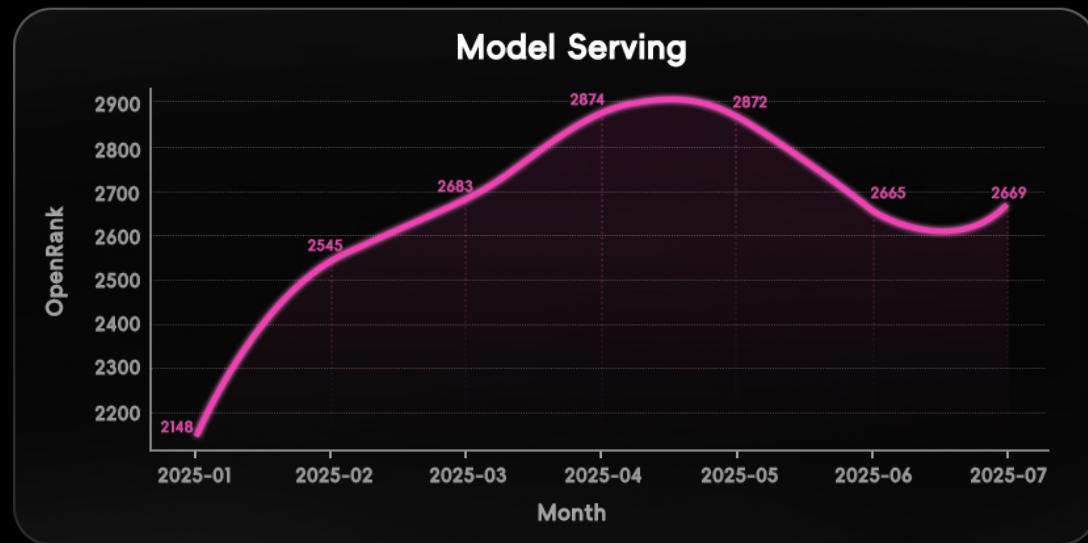
根据用户所在组织的规模做分段，提出了一种双许可限制，不同规模组织下的用户使用不同的许可：1. 如果是个人用户或者所在组织是 10 人及以下，采用 AGPLv3，这也是一种 copyleft 协议，用户可以免费使用，但如果做了修改和分发，必须同样开源并提供完整的源代码；2. 超过 10 人的情况，则需要联系 Cherry Studio 的团队进行商业授权。

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限制属性，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。

技术领域的趋势 - Agent



技术领域的趋势 - Infra



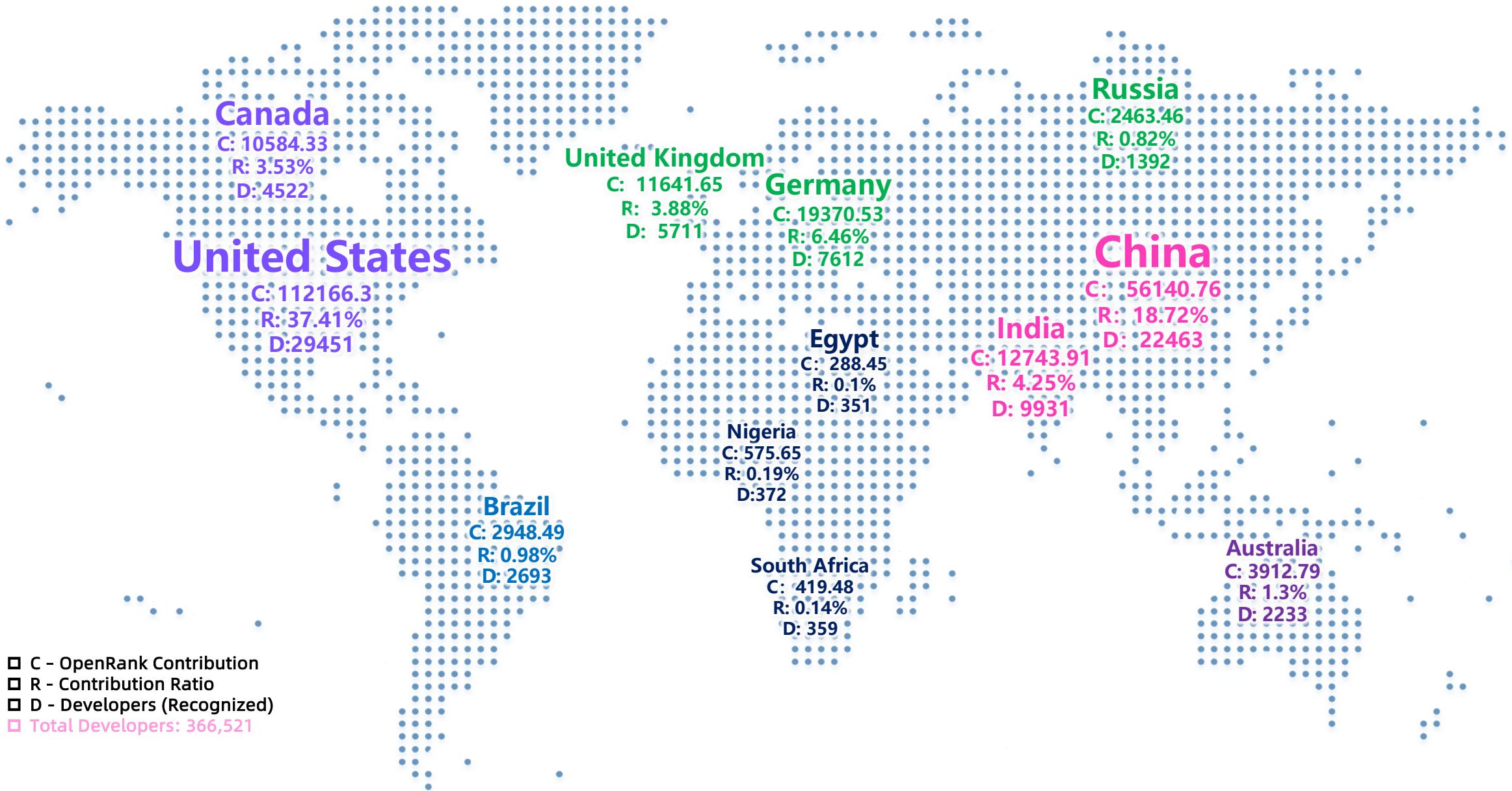
边缘地带的项目们



Project	OpenRank	Star	OpenRank Trend	Language	Created	Comment
_invokeai	48	25,630		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 Star 数量和更为陡峭的 OpenRank 下降曲线。
onyx	46	13,254		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。
DeerFlow	37	15,954		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。
Mooncake	36	3,704		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。
Transformers	34	14,782		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。
COSYVOICE	32	15,548		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。
Agent2Agent Protocol	29	18,825		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。

注：以上数据截止 2025 年 8 月 1 日

大模型生态下全球开发者分布画像



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%

大模型生态整体以中美开发者的贡献为主导：

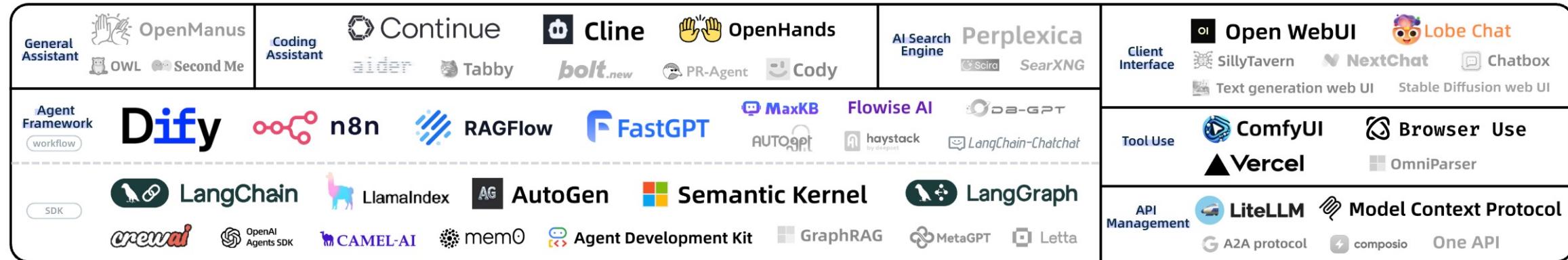
- 在 **AI Infra** 领域中美的领先地位更加明显，两国在基础设施领域的贡献度达到 60% 以上；
- AI Data** 领域全球的参与情况更加平均，中美的总体贡献占比仅 46.5%，欧洲各国如波兰、挪威、法国、荷兰等国的参与度均进入全球前十；
- AI Agent** 领域中美差距大幅缩小，贡献度占比分别为 24.6% 和 21.5%，中国开发者在 Agent 层面相较其他领域的投入更多。

注：开发者数量基于 GitHub 首页信息统计；贡献度和贡献比例也使用 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](#)

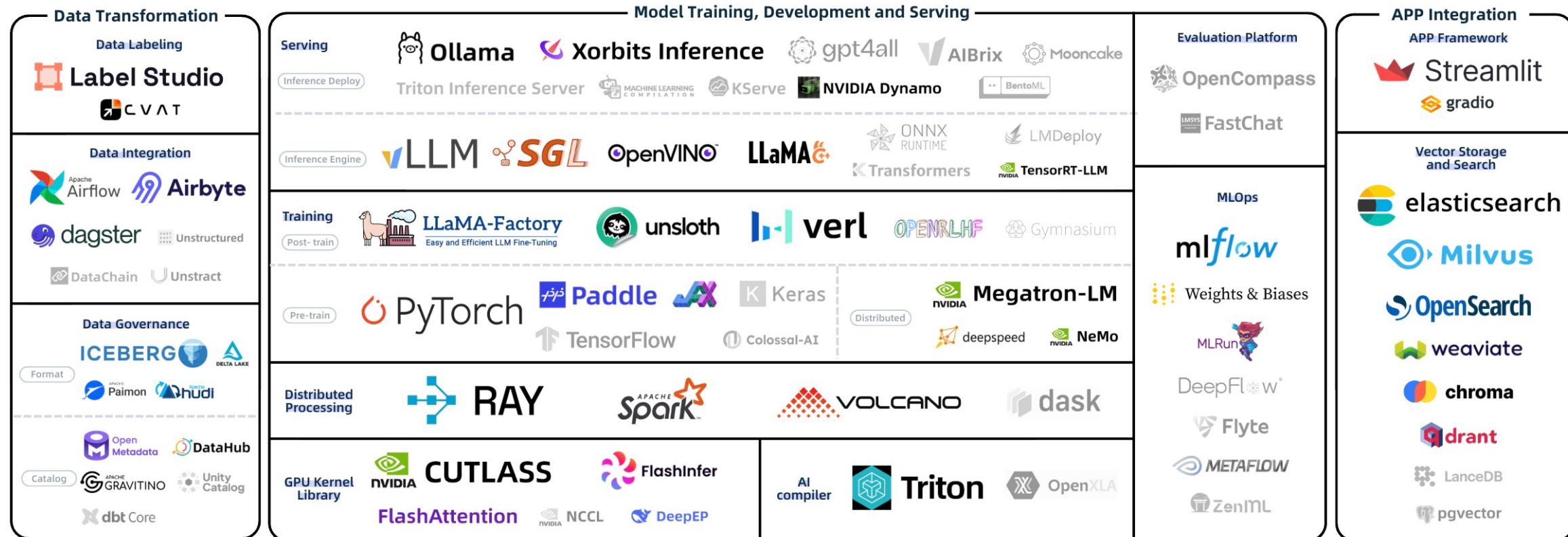
100 天之后，全景图的变与不变

哪些领域和项目出局了？

Application



Infrastructure



出局的项目中，有不少可能正在步入“AI 墓园”的路上



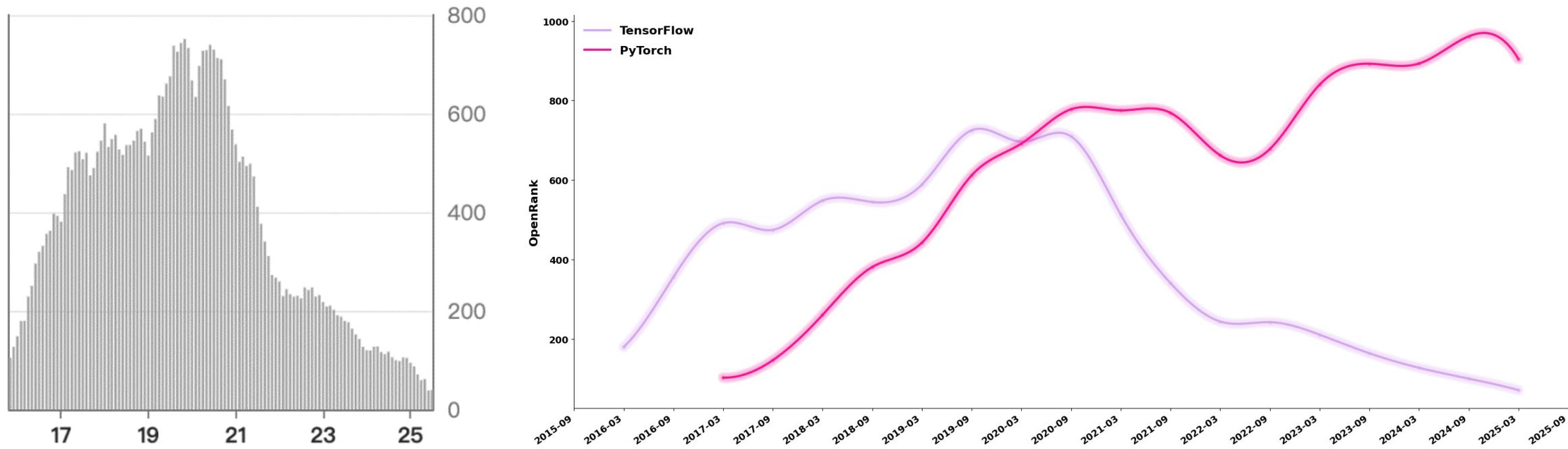
AI Agent			
Project	Domain	Star	Trend
OpenManus	General Agent	49,279	
OWL	General Agent	17,936	
NextChat	Chatbot	85,665	
bolt.new	AI Coding	15,584	

AI Infra			
Project	Domain	Star	Trend
MACHINE LEARNING COMPILATION	Inference Deploy	21,188	
gpt4all	Inference Deploy	76,557	
FastGPT	Inference Enginee	39,036	
Text Generation Inference	Inference Enginee	10,453	

- 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



- **2015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“Pythonic”和“研究人员优先”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代码和学习新 API 的负担, 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。**

哪些领域和项目第一次进入视野



Agentic AI

The banner is organized into several sections:

- AI Coding:** GEMINI, OPENCODE, Cline, codename goose, Continue, OpenHands, marimo, Codex CI, avante.nvim.
- Chatbot & Knowledge Management:** Cherry Studio, Open WebUI, Lobe Chat.
- Embodied Agent:** GENESIS, xiaozhi-esp32.
- Agent Workflow Platform:** Dify, n8n, RAGFlow, Langflow, Mastra, activepieces, MaxKB, FastGPT, Flowise AI.
- Agent Tool:** LiteLLM, supabase, Vercel, ComfyUI, mem0, Agent Development Kit, Browser Use, Model Context Protocol.
- Agent Framework:** LangGraph, Pydantic AI, LangChain, springAI, LiveKit Agents, Llamaindex, Semantic Kernel, Pipecat, AutoGen, gno, CAMEL-AI, OpenAI Agents SDK, ELIZA OS, crewai.

AI Infra

AI Infra

Model Training, Development and Serving

Serving	Ollama	NVIDIA Dynamo	Xorbits Inference	ramalama	GPUStack		
Inference Engine	LLM	SGL	NVIDIA TensorRT-LLM	OpenVINO™	LLaMA		
Training	swift	unsloth	LLaMA-Factory	Reinforce Learning	verl	OPENRLHF	AReAL
Fine-tune							
Training Platform	PyTorch	Paddle	Distributed Training	Megatron-LM	deepspeed	NVIDIA NeMo	
				nvidia		nvidia	
Distributed Compute	RAY	APACHE Spark	VOLCANO				
AI Kernel Library	RAPIDS	NVIDIA TransformerEngine	FlashInfer	AI Compiler	Triton	Modular	
FlashAttention	MLX	CUTLASS	DeepEP				

LLMops

mlflow
1Panel
Langfuse
Weights & Biases
opik
Phoenix
MLRun
promptfoo
Dagger

AI Data

Data Labeling	Label Studio
c v a t	Vespa
Data Integration	Apache Airflow
	Airbyte
	dagster
Vector Storage and Search	elasticsearch
	Milvus
	OpenSearch
	chroma
	weaviate
	drant
APP Framework	Streamlit
	gradio

新进项目中最活跃的开源项目 Top 10



#	Project	Domain	OpenRank	Stars	OpenRank Trend	Language	Created	Initiator	License
1	GEMINI	AI Coding	391	66,881		TypeScript	2025-04-17	Google	Apache 2.0
2	Cherry Studio	Chatbot	313	30,976		TypeScript	2024-05-24	Cherry Studio	User-Segmented Dual Licensing
3	OPENCODE	AI Coding	195	17,071		Go	2025-04-30	Anomaly Innovations Inc	MIT
4	supabase	Agent Tool	178	86,533		TypeScript	2019-10-12	Supabase	Apache 2.0
5	Langflow	Agent Platform	143	95,122		Python	2023-02-08	IBM DataStax (Acquired)	MIT
6	codename goose	AI Coding	139	18,122		Rust	2024-08-23	Block Inc	Apache 2.0
7	Mastra	Agent Platform	135	15,510		TypeScript	2024-08-06	Mastra	Apache 2.0
8	swift <small>LLM training and deployment framework</small>	Fine-tune	113	9,063		Python	2023-08-01	AlibabaCloud	Apache 2.0
9	Agno <small>Reasoning Agents</small>	Multi-agent Framework	106	31,212		Python	2022-05-04	Agno	MPL 2.0
10	Modular	AI Compiler	98	24,590		Mojo	2023-04-28	Modular	Apache 2.0

注：以上数据截止 2025 年 8 月 1 日

没变的是：此消彼长，前浪后浪，增长与衰落，一如既往



AI Agent

<p>AI Coding</p> GEMINI <small>2025 NEW</small> OPENCODE <small>2025 NEW</small> Cline <small>codename goose</small> Continue OpenHands marimo Codex CLI avante.nvim	<p>Chatbot & Knowledge Management</p> Cherry Studio LibreChat AstrBot siyuan Docling Anything LLM	<p>Embodied Agent</p> GENESIS xiaozi-esp32
<p>Agent Workflow Platform</p> Dify n8n RAGFlow Langflow Mastra activepieces MaxKB FastGPT Flowise AI	<p>Agent Tool</p> LiteLLM supabase Vercel ComfyUI mem0 Agent Development Kit Browser Use Model Context Protocol	
<p>Agent Framework</p> LangGraph Pydantic AI LangChain spring AI Llamaindex Semantic Kernel Pipecat AutoGen	<p>Multi-agent Framework</p> LiveKit Agents Agno CAMEL-AI OpenAI Agents SDK ELIZA OS crewai	

AI Infra

Model Training, Development and Serving								AI Data								
<p>Serving</p> Ollama NVIDIA Dynamo <small>2025 NEW</small> Xorbits Inference ramalama GPUStack LLM SGL TensorRT-LLM OpenVINO LLaMA						<p>LLMops</p> mlflow 1Panel Langfuse Weights & Biases opik Phoenix MLRun promptfoo Dagger										
<p>Training</p> swift unsloth LLaMA-Factory PyTorch Paddle						<p>Data Labeling</p> Label Studio CVAT Vespa										
verl OPENRLHF AReal Megatron-LM deepspeed NeMo						<p>Data Integration</p> Apache Airflow Airbyte dagster										
RAY APACHE Spark VOLCANO						<p>Data Governance</p> ICEBERG APACHE Paimon DataHub DELTA LAKE Open Metadata APACHE GRAVITINO Hudi										
RAPIDS TransformerEngine FlashInfer DeepEP MLX CUTLASS						<p>Vector Storage and Search</p> elasticsearch Milvus OpenSearch chroma weaviate drant										

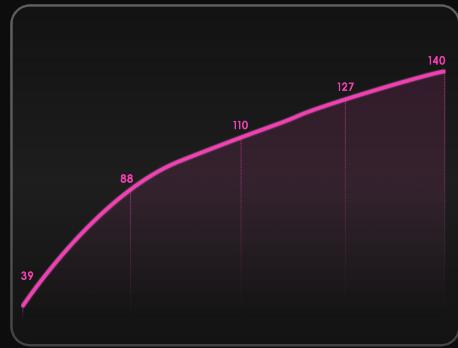
全景图上的「The New Wave」



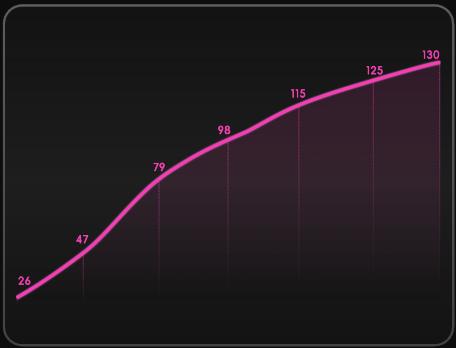
Project	Domain	OpenRank	Stars	Language	Created	Initiator	License
OPENCODE	AI Coding	195	17,071	Go	2025-04-30	Anomaly Innovations Inc	MIT
GEMINI	AI Coding	391	66,881	TypeScript	2025-04-17	Google	Apache 2.0
Codex CLI	AI Coding	81	35,282	Rust	2025-04-13	OpenAI	Apache 2.0
Agent Development Kit	Agent Tool	109	11,477	Python	2025-04-01	Google	Apache 2.0
OpenAI Agents SDK	Multi-agent Framework	72	13,225	Python	2025-03-11	OpenAI	MIT
NVIDIA Dynamo	Inference Deploy	140	4,642	Rust	2025-03-03	NVIDIA	Apache 2.0

注：以上数据截止 2025 年 8 月 1 日

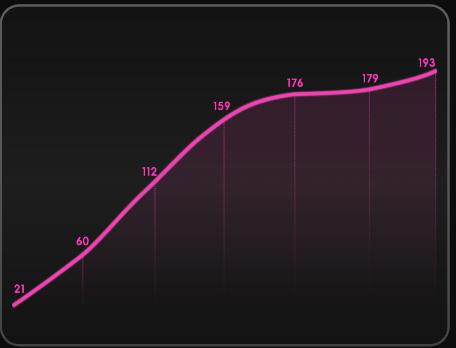
全景图上的「Up and Down」



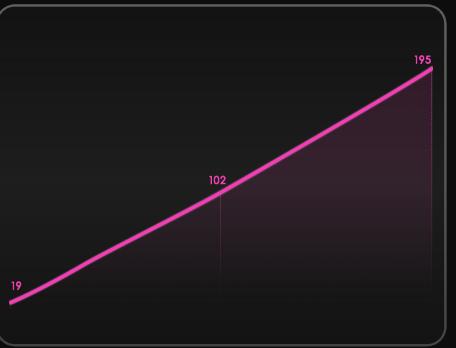
ai-dynamo/dynamo (+101)



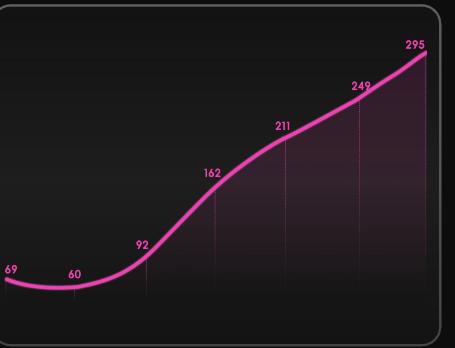
mastra-ai/mastra (+109)



volcengine/verl (+172)



sst/opencode (+176)



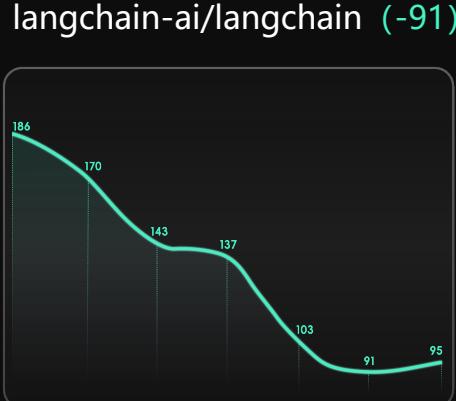
NVIDIA/TensorRT-LLM (+226)

Decrease

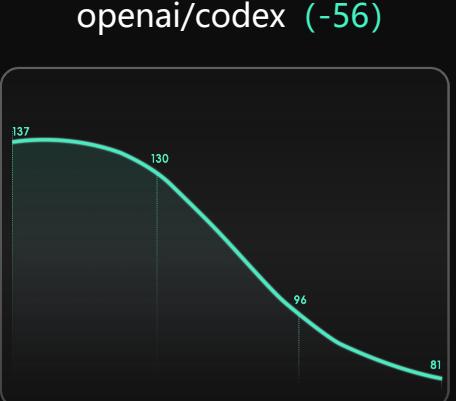
(OpenRank Delta) Growth



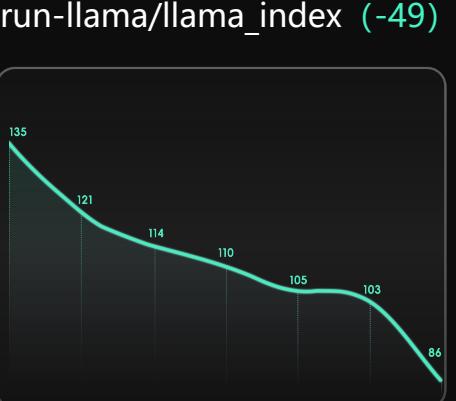
elizaOS/eliza (-233)



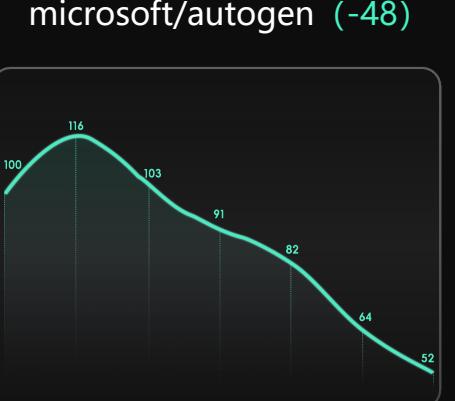
langchain-ai/langchain (-91)



openai/codex (-56)



run-llama/llama_index (-49)

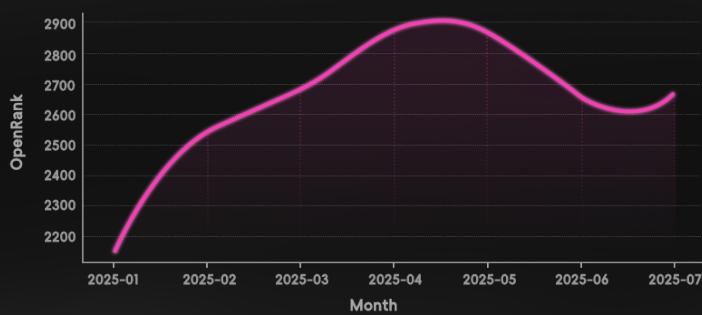


microsoft/autogen (-48)

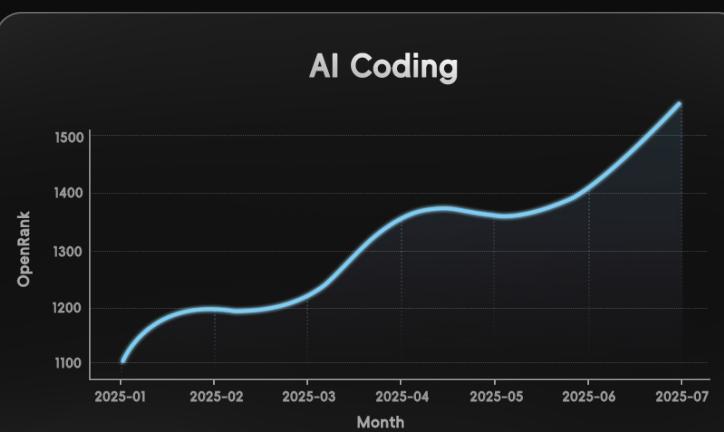
专题洞察

技术趋势下的项目故事

Model Serving



AI Coding



AI Agent



Model Serving

Serving



NVIDIA Dynamo

Xorbits Inference

ramalama

GPUSTack

Inference Engine

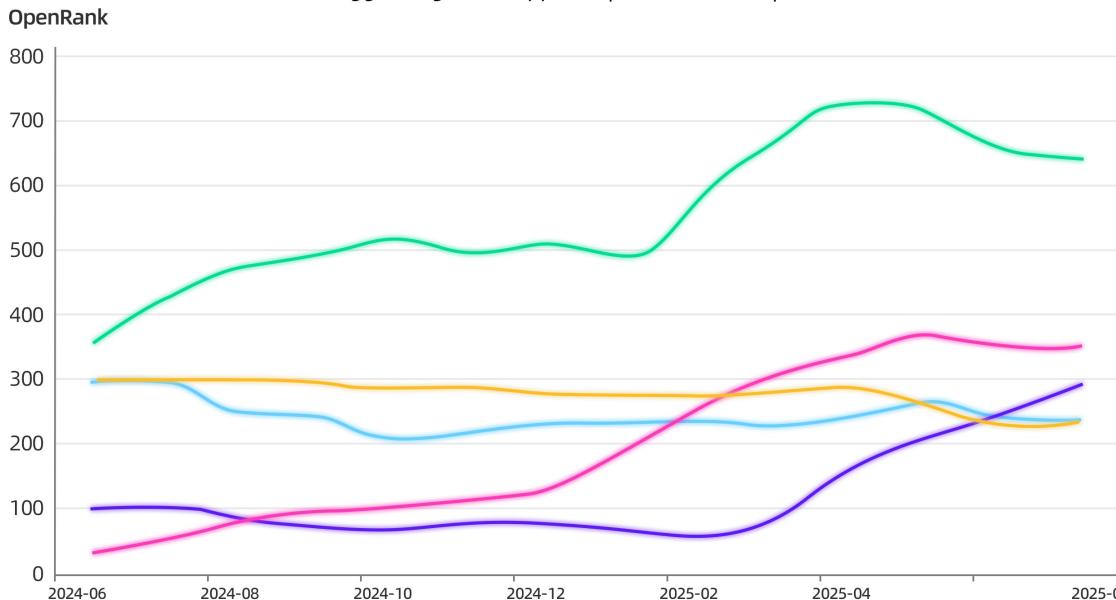


TensorRT-LLM

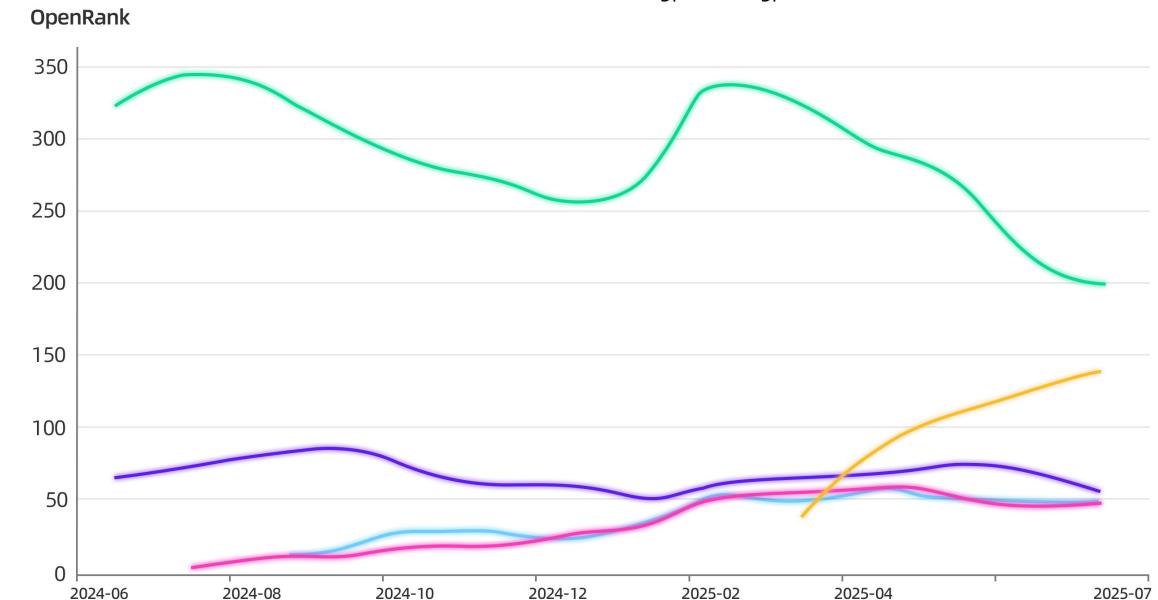
OpenVINO™

LLaMAČ+

vllm-project/vllm sgl-project/sclang NVIDIA/TensorRT-LLM
ggml-org/llama.cpp openvinotoolkit/openvino



ollama/ollama ai-dynamo/dynamo xorbitsai/inference
containers/ramalama gpustack/gpustack



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。

集群推理： vLLM SGL

- 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。

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。

AI Coding

AI Coding



Cline



codename
goose



OpenHands



marimo



Codex CLI



avante.nvim

CLI 形态

Open Source:    Codex CLI

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:  marimo

Closed Source:  CURSOR  Windsurf  Tera  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



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。

模型基座

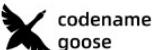
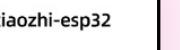
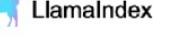
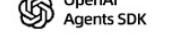
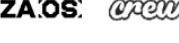
 Claude

 Gemini

 OpenAI

 Qwen3-Coder

AI Agent

AI Coding         	Chatbot & Knowledge Management        	Embodied Agent  
Agent Workflow Platform         	Agent Tool        	
Agent Framework        	LiveKit Agents     	

记忆：memo

- 技术定位：**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。
- 开源社区：**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。

工具：Browser Use

- 技术定位：**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。
- 开源社区：**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。

执行（工作流）：Dify

- 技术定位：**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。
- 开源社区：**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。

交互：Lobe Chat

- 技术定位：**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。
- 开源社区：**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。

规划：LLM + Prompt

Thanks

🌐 全景图地址: <https://antoss-landscape.my.canva.site/>

⌚ 开源地址: <https://github.com/antgroup/llm-oss-landscape>

关注**蚂蚁开源公众号**, 获得 Landscape 解读文章

敬请期待我们围绕着技术趋势的项目故事系列解读

