2023 中国开源开发者报告

China Open Source 2023 Annual Report

LLM 技术报告

出品：OSCHINA & Gitee

编委会：

雨多田光，OSCHINA 总编
局长，OSCHINA 主编
王茜，OSCHINA 主编
叶子，OSCHINA 新媒体运营
鱼仔，OSCHINA 新媒体运营
诺墨，Gitee 开源社区产品负责人
张力文，Gitee 公有云研发负责人
李泽辰，Gitee 主编
李涛，APUS 董事长兼 CEO

设计：张琪

2023 年 12 月发布

LLM 技术报告

大语言模型（LLM） 技术作为人工智能领域的一项重要创新在今年引起了广泛的关注。

LLM 是利用深度学习和大数据训练的人工智能系统，专门设计来理解、生成和回应自然语言。这些模型通过分析大量的文本数据来学习语言的结构和用法，从而能够执行各种语言相关任务。以 GPT 系列为代表，LLM 以其在自然语言处理领域的卓越表现，成为推动语言理解、生成和应用的引擎。

LLM 在多个领域都取得了令人瞩目的成就。在自然语言处理领域，GPT 系列模型在文本生成、问答系统和对话生成等任务中展现出色的性能。在知识图谱构建、智能助手开发等方面，LLM 技术也发挥了关键作用。此外，它还在代码生成、文本摘要、翻译等任务中展现了强大的通用性。

本报告从技术人视角出发，将深入探讨 LLM 技术的背景、基础设施、应用现状，以及相关的工具和平台。

LLM Tech Map

大模型

- 备案上线的中国大模型
- 知名大模型
- 知名大模型应用

AI 编程

- 插件、IDE、终端
- 代码生成工具

工具和平台

- LLMOps
- 大模型聚合平台
- 开发工具

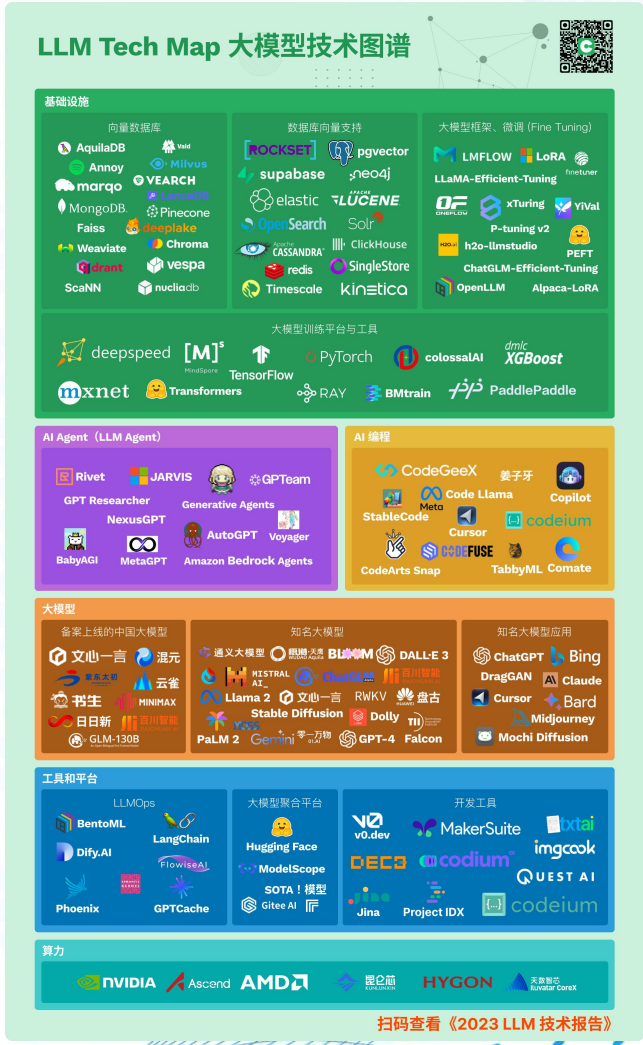
基础设施

- 向量数据库
- 数据库向量支持
- 大模型框架、微调 (Fine Tuning)
- 大模型训练平台与工具

算力

LLM Agent

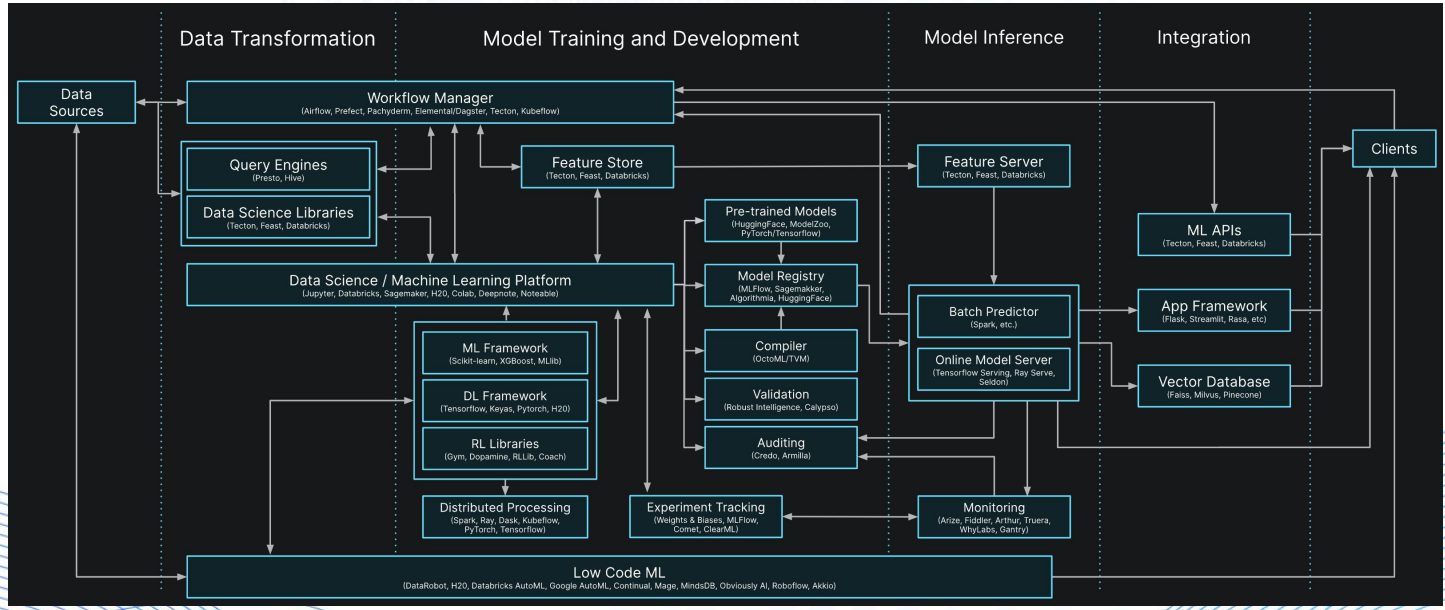
编程语言



LLM 技术背景

Transformer 架构和预训练与微调策略是 LLM 技术的核心，随着大规模语言数据集的可用性和计算能力的提升，研究者们开始设计更大规模的神经网络，以提高对语言复杂性的理解。

GPT (Generative Pre-trained Transformer) 的提出标志着 LLM 技术的飞速发展，其预训练和微调的方法为语言任务提供了前所未有的性能，以此为基础，多模态融合的应用使得 LLM 更全面地处理各种信息，支持更广泛的应用领域。



图源：https://postgresml.org/docs/.gitbook/assets/ml_system.svg

LLM 基础设施

01

向量数据库/数据库向量支持

为大模型提供高效的存储和检索能力

02

大模型框架及微调 (Fine Tuning)

大模型框架提供基本能力和普适性，而微调则是实现特定应用和优化性能的关键环节

03

大模型训练平台&工具

提供了在不同硬件和环境中训练大语言模型所需的基础设施和支持

04

编程语言

以 Python 为代表

LLM 基础设施：向量数据库/数据库向量支持

向量数据库是专门用于存储和检索向量数据的数据库，它可以为 LLM 提供高效的存储和检索能力。通过数据向量化，实现了在向量数据库中进行高效的相似性计算和查询。

根据向量数据库的的实现方式,可以将向量数据库大致分为两类：

➢ 原生向量数据库

原生的向量数据库专门为存储和检索向量而设计，所管理的数据是基于对象或数据点的向量表示进行组织和索引。
包括 Chroma、LanceDB、Margo、Milvus、Pinecone 等均属于原生向量数据库。

➢ 添加“向量支持”的传统数据库

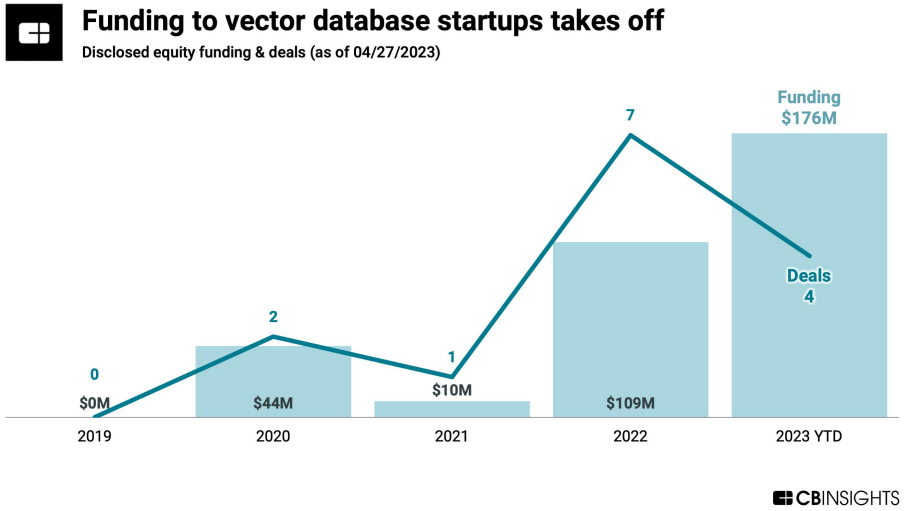
除了选择专业的向量数据库，对传统数据库添加“向量支持”也是主流方案。比如 Redis、PostgreSQL、ClickHouse、Elasticsearch 等传统数据库均已支持向量检索。

LLM 基础设施：向量数据库/数据库向量支持

自 2022 年 ChatGPT 问世以来，大模型星火初始，向量数据库不但获得了技术领域的关注，也逐渐吸引了市场和资本的注意力。近两年来，向量数据库公司迎来了一波融资潮：

- Pinecone：已融资 1.38 亿美元
- Zilliz：已融资 1.15 亿美元
- Weaviate：已融资 6770 万美元
- Vespa：已融资 3100 万美元
- Chroma：已融资 2000 万美元
- Qdrant：已融资 980 万美元
- Marqo：已融资 440 万美元
- LanceDB：已融资 50 万美元
- ……

据西南证券研究发展中心预测，2025 年向量数据库渗透率约为 30%，则全球向量数据库市场规模约为 99.5 亿美元，中国向量数据库市场规模约为 82.56 亿元。



*2023 年前四个月，向量数据库公司融资额**已经达到1.76亿美元**，超过了 2022 年的总和*

（图源：https://www.cbinsights.com/research/generative-ai-infrastructure-vector-database/）

LLM 基础设施：大模型框架及微调 (Fine Tuning)

大模型框架指专门设计用于构建、训练和部署大型机器学习模型和深度学习模型的软件框架。这些框架提供了必要的工具和库，使开发者能够更容易地处理大量的数据、管理巨大的网络参数量，并有效地利用硬件资源。

微调（Fine Tuning）是在大模型框架基础上进行的一个关键步骤。在模型经过初步的大规模预训练后，微调是用较小、特定领域的数据集对模型进行后续训练，以使其更好地适应特定的任务或应用场景。这一步骤使得通用的大型模型能够在特定任务上表现出更高的精度和更好的效果。

大模型框架提供了 LLM 的基本能力和普适性，而微调则是实现特定应用和优化性能的关键环节。两者相结合，使得 LLM 在广泛的应用场景中都能发挥出色的性能。

LLM 基础设施：大模型框架及微调 (Fine Tuning)

大模型框架有哪些特点：

- **抽象和简化**：大模型开发框架通过提供高层次的 API 简化了复杂模型的构建过程。这些 API 抽象掉了许多底层细节，使开发者能够专注于模型的设计和训练策略。
- **性能优化**：这些框架经过优化，以充分利用 GPU、TPU 等高性能计算硬件，以加速模型的训练和推理过程。
- **易于扩展**：为了处理大型数据集和大规模参数网络，这些框架通常设计得易于水平扩展，支持在多个处理器或多个服务器上并行处理。
- **支持大数据集**：它们提供工具来有效地加载、处理和迭代大型数据集，这对于训练大型模型尤为重要。

国产深度学习框架 OneFlow 架构
（图源：https://www.oneflow.org/a/chanpin/oneflow/）

LLM 基础设施：大模型框架及微调 (Fine Tuning)

想要微调一个模型，一般包含以下关键步骤：

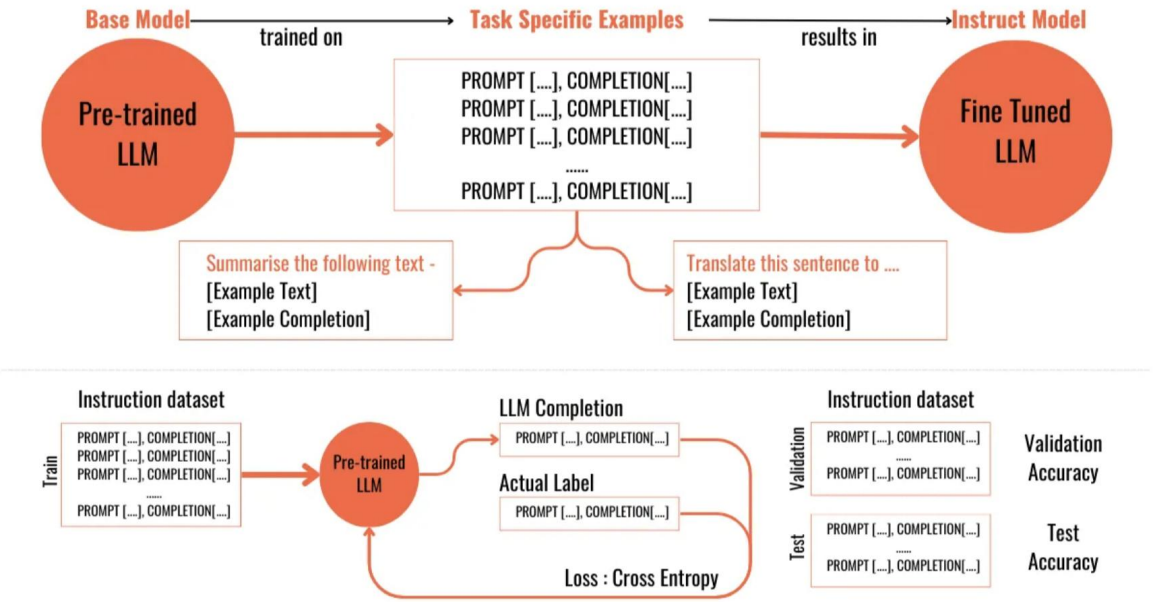
1.选择预训练模型：选取一个已经在大量数据上进行过预训练的模型作为起点；

2.准备任务特定数据：收集与目标任务直接相关的数据集，这些数据将用于微调模型；

3.微调训练：在任务特定数据上训练预训练的模型，调整模型参数以适应特定任务；

4.评估：在验证集上评估模型性能，确保模型对新数据有良好的泛化能力；

5.部署：将性能经验证的模型部署到实际应用中去。



微调的过程也是分类模型训练的过程

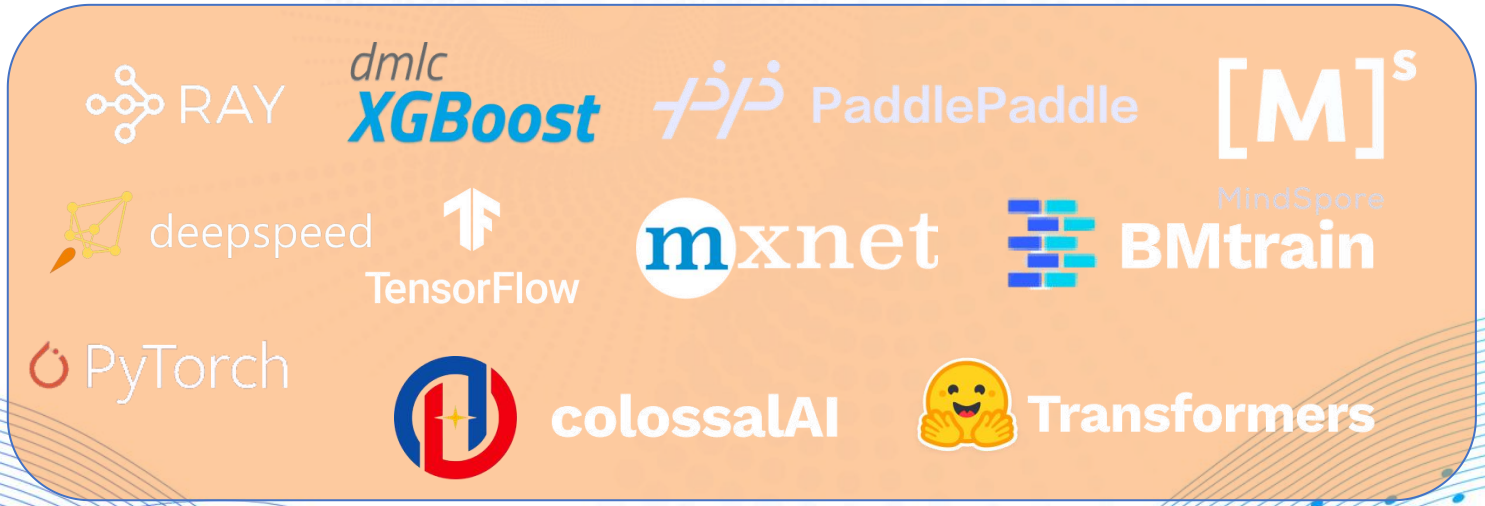
(图源：https://medium.com/mlearning-ai/what-is-a-fine-tuned-llm-67bf0b5df081)

LLM 基础设施：大模型训练平台与工具

大模型训练平台和工具提供了强大且灵活的基础设施，使得开发和训练复杂的语言模型变得可行且高效。

这些工具提供了先进的算法、预训练模型和优化技术，极大地简化了模型开发过程，加速了实验周期，并使得模型能够更好地适应各种不同的应用场景。此外，它们还促进了学术界和工业界之间的合作与知识共享，推动了自然语言处理技术的快速发展和广泛应用。

相比前边的大模型框架和微调，一言以蔽之：**平台化（训练平台）、灵活化（各种工具）**



LLM 基础设施：大模型训练平台与工具

大模型训练平台与工具根据其性质不同，可分为以下几类：

1.云服务和商业平台

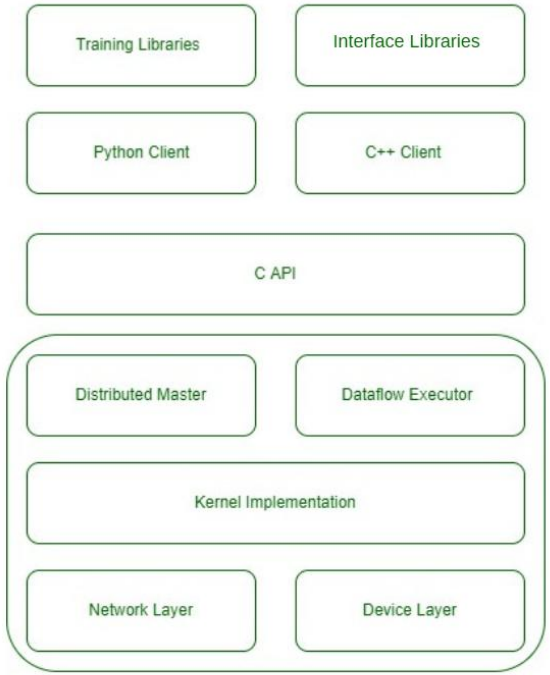
这些平台提供了从模型开发到部署的综合解决方案，包括计算资源、数据存储、模型训练和部署服务。它们通常提供易于使用的界面，支持快速迭代和大规模部署。Amazon SageMaker、Google Cloud AI Platform 和 Microsoft Azure Machine Learning 都是提供端到端机器学习服务的云平台。

2.专用硬件加速工具

这些工具和库专门为加速机器学习模型的训练和推理而设计，通常利用 GPU 或 TPU 等硬件。这类工具可以显著提高训练和推理的速度，使得处理大规模数据集和复杂模型变得可行。NVIDIA CUDA 和 Google Cloud TPU 均是此类工具。

3.开源框架和库

这类工具通常由开源社区支持和维护，提供了灵活、可扩展的工具和库来构建和训练大型机器学习模型，如 TensorFlow 和 PyTorch 和 Hugging Face Transformers 等。



TensorFlow 架构图
(图源：https://www.geeksforgeeks.org/architecture-of-tensorflow/)

LLM 基础设施：编程语言

LLM 的训练和应用通常使用多种编程语言，取决于任务的需求和团队的偏好。

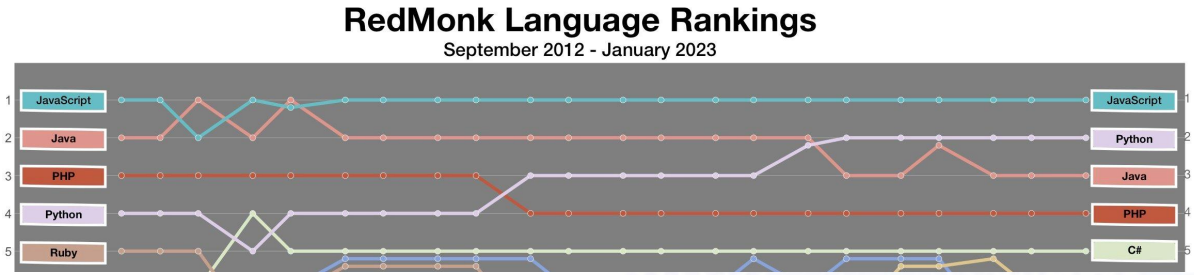
Python 是 LLM 开发中最常用的编程语言。它的广泛使用得益于其简洁的语法、强大的库支持（如 NumPy, Pandas, Matplotlib）和深度学习框架（如TensorFlow, PyTorch, Keras）。

此外，**AI 开发领域也有崛起的新秀语言 Mojo** ，C++ 有时用于优化计算密集型任务，而 Java 在企业环境中处理模型部署和系统集成方面常见。JavaScript 适用于 Web 环境的 LLM 应用。

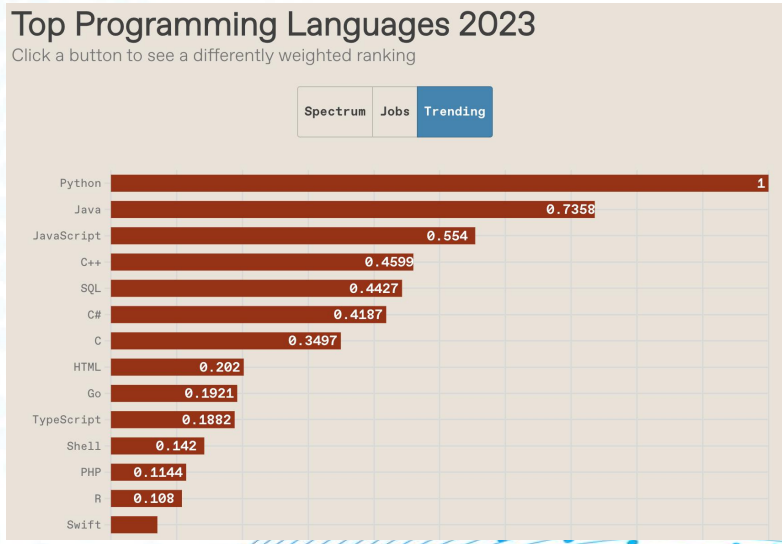
LLM 基础设施：编程语言

2023 年是大语言模型 (LLM) 之年，Python 作为人工智能领域使用度最高的编程语言，在 2023 年到底有多火？

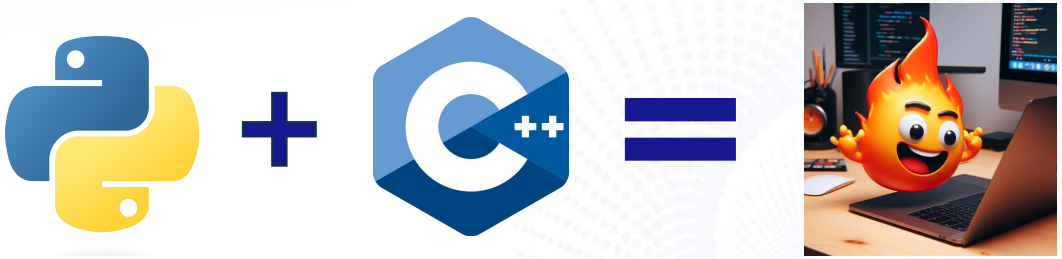
从各种开发者报告、编程语言榜单来看。只要出现有关编程语言流行度的排名，Python 始终位列前茅，而 Java、C/C++ 等同样在 LLM 开发中发挥关键作用的语言紧随其后。



Dec 2023	Dec 2022	Change	Programming Language	Ratings	Change
1	1		Python	13.86%	-2.80%
2	2		C	11.44%	-5.12%
3	3		C++	10.01%	-1.92%
4	4		Java	7.99%	-3.83%
5	**5**		**C#**	**7.30%**	**+2.38%**



LLM 基础设施：编程语言



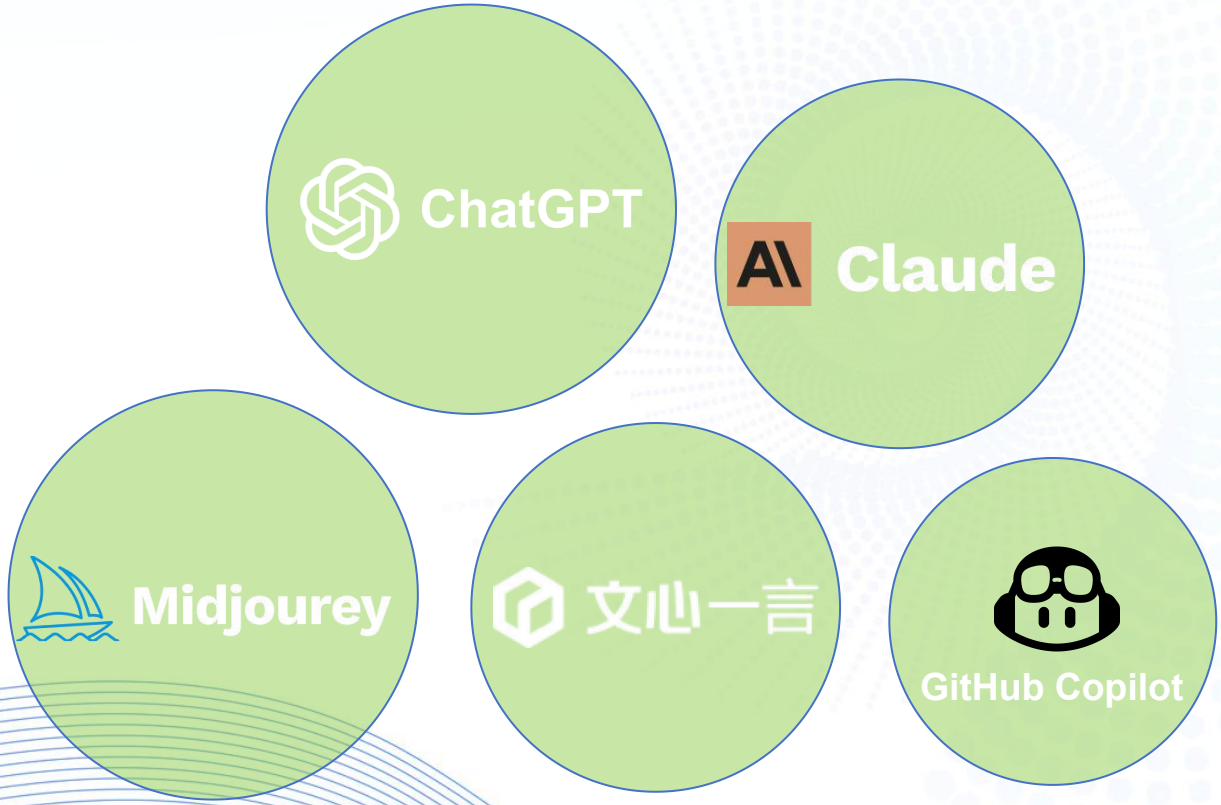
LANGUAGES	TIME (S) *	SPEEDUP VS PYTHON
PYTHON 3.10.9	1027 s	1x
PYPY	46.1 s	22x
SCALAR C++	0.20 s	5000x
MOJO 🔥	0.03 s	68000x

新势力 Mojo 🔥

- 2023 年 9 月面向大众开放
- **LLVM/Swift 之父**创业公司 Modular AI 开发
- 结合了 Python 的易用性以及 C 语言的**可移植性和性能**
- 支持与任意 **Python** 代码**无缝集成**
- 性能是 Python 的 **68000** 倍

Mojo 与其他语言性能对比
（图源：https://www.modular.com/max/mojo）

大模型应用现状



2022 年底大模型应用 ChatGPT 发布后，点燃了世界范围内对于大模型技术及其应用的关注和热情。2023 年，国内外各大厂商均投身于大模型的浪潮当中，涌现了诸多知名的大模型及应用，它们结合了文本、图片、视频、音频多种介质，在文本生成、图片生成、AI 编程等方向均有出色的表现。

大模型应用现状：知名大模型

在全球范围内，已经发布了多款知名大模型，这些大模型在各个领域都取得了突破性的进展。

处理文本数据的 GPT-4，能同时处理和理解多种类型数据的多模态模型 DALL-E 3，以及开源大模型的代表 Llama 2 都在短时间内获得了大量关注和用户，构成了大模型领域的「第一梯队」。

大模型应用现状：首批备案上线的中国大模型

8 月 31 日，百度、字节、商汤、中科院旗下紫东太初、百川智能、智谱华章等 8 家企业 / 机构的大模型产品首批通过《生成式人工智能服务管理暂行办法》备案，可正式上线面向公众提供服务。

具体包括：百度（文心一言）、抖音（云雀大模型）、智谱 AI（GLM 大模型）、中科院（紫东太初大模型）、百川智能（百川大模型）、商汤（日日新大模型）、MiniMax（ABAB 大模型）、上海人工智能实验室（书生通用大模型）、腾讯（混元大模型，9月15日通过）。

大模型应用现状：知名大模型应用

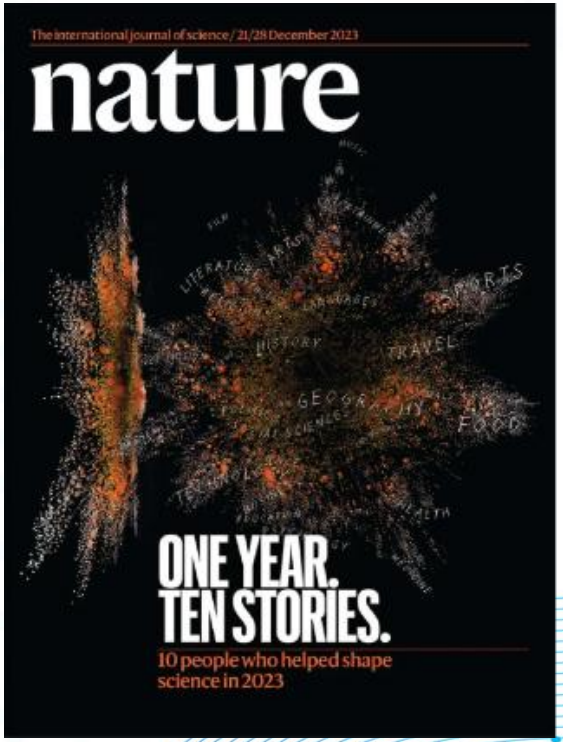
LLM 已经在多种应用场景中得到了应用，包括文本生成、机器翻译、问答、自然语言推理等。

以 ChatGPT、Claude、 Bard 为代表的文本生成应用，Midjourney 为代表的图片生成应用，以 GitHub Copilot 为代表的 AI 编程应用，以 HeyGen 为代表的数字人生成应用，在推出后都获得了大量用户的青睐。

ChatGPT、Ilya Sutskever

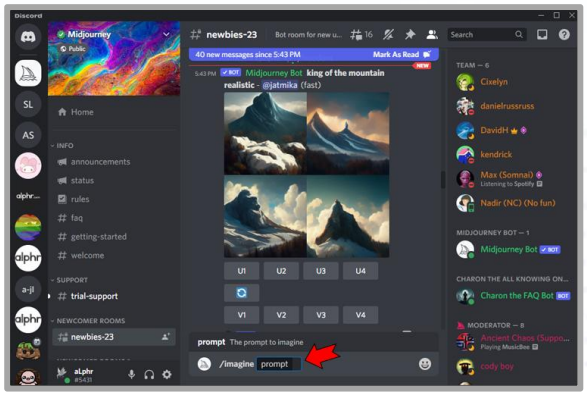
入选《Nature》
年度十大科学人物

OpenAI 联合创始人和前首席科学家 Ilya Sutskever

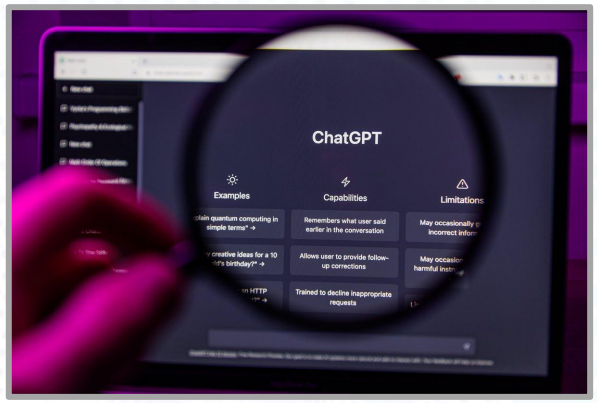


《自然》2023 年度十大人物中，ChatGPT 破例成为第 11 人
（图源：https://www.nature.com/articles/d41586-023-03930-6）

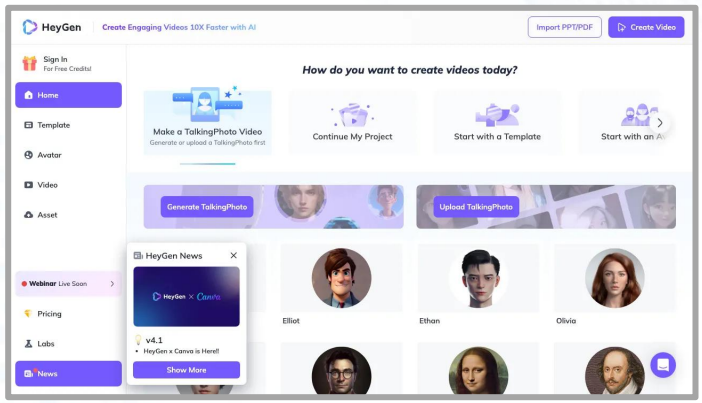
大模型应用现状：知名大模型应用



Midjourney：最强文本生成图像 AI 应用



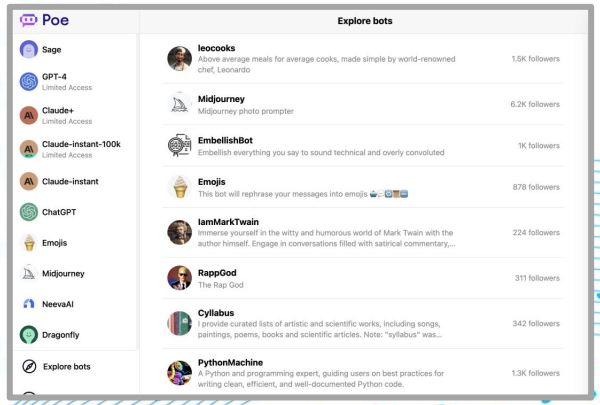
ChatGPT：最强生成式对话大模型产品



HeyGen：年度最火热 AI 视频生成工具

APUS：千亿级多模态通用 AI 大模型

Codeverter：代码转换器

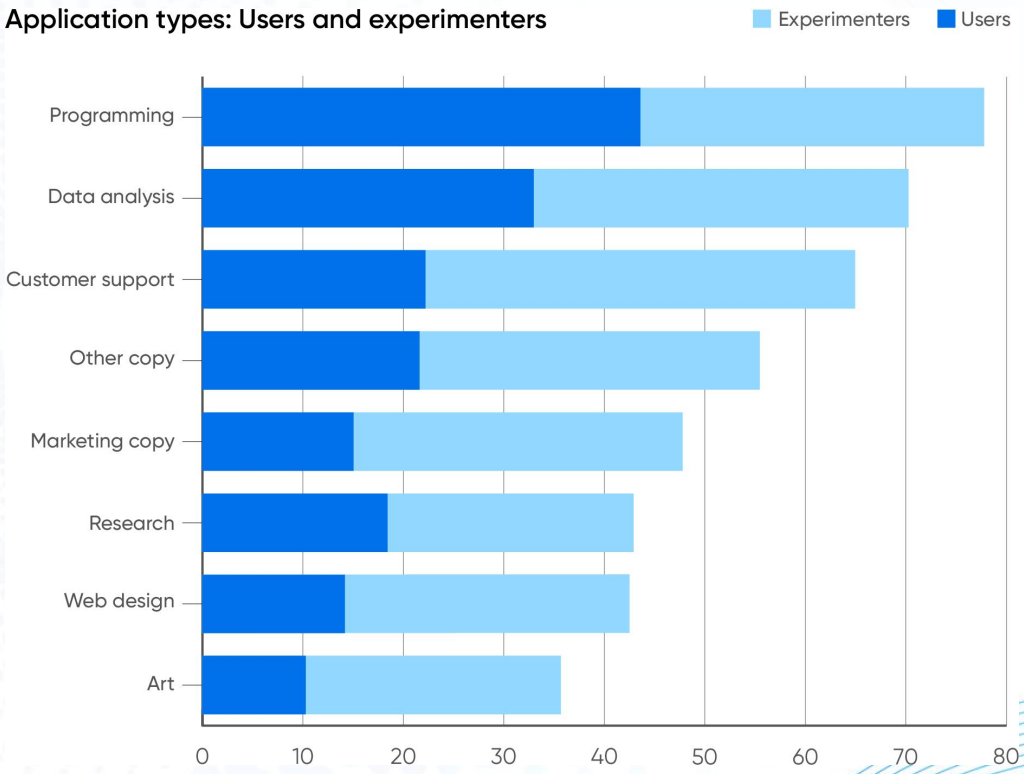


Poe：集成主流大模型的对话机器人

AI 编程

生成式 AI 正经历前所未有的快速普及，而开发者们正积极将 AI 作为自己的生产力工具，随着众多 AI 编程工具的普及，开发者们使用 AI 辅助工作已经逐渐司空见惯。

分析公司 O’Reilly 日前发布一份《2023 Generative AI in the Enterprise》报告， 报告中指出，**有 77% 受访者正在使用 AI 来辅助编程**。



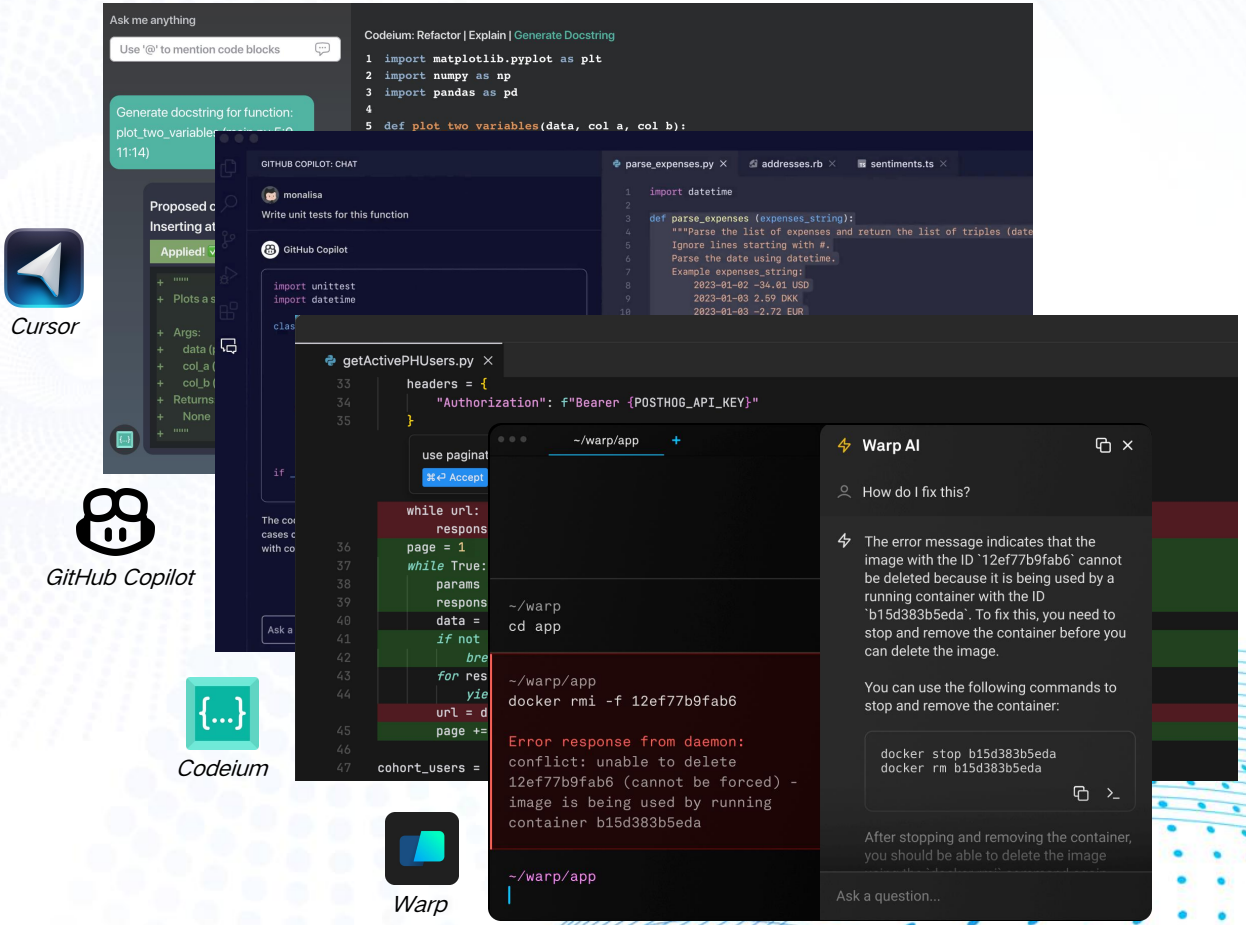
图源：https://www.oreilly.com/radar/generative-ai-in-the-enterprise/

AI 编程工具：插件、IDE、终端

目前最常见的 AI 编程工具大多以插件、IDE 和终端的形式出现，它们大多交互直观且使用门槛低，大大提高了 AI 编程工具的使用率。

GitHub Copilot 和 Codeium 是比较常见的 AI 编程插件，而 Cursor 和 Warp 分别是具有 AI 编程能力的 IDE 和终端工具。

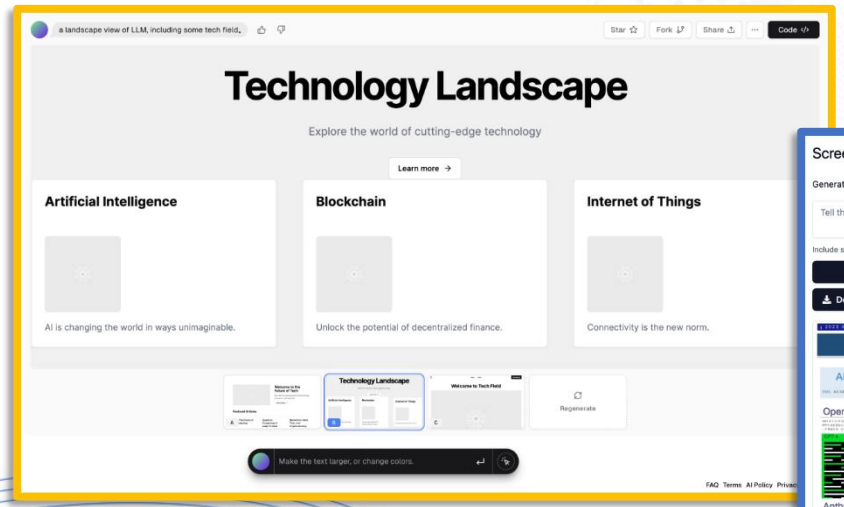
除了海外产品，国内如姜子牙、CodeFuse、CodeGeeX、百度 Comate 等都是十分优秀的 AI 编程工具。



AI 编程新形态：代码生成工具

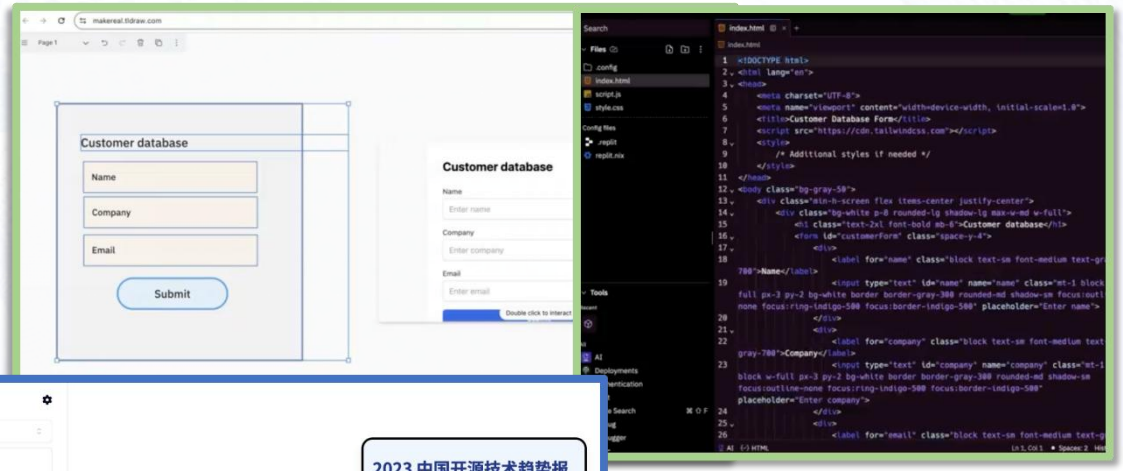
通过原型或图片直接生成包含代码的完整页面，**生成的样式一般配有多种可选语言的代码**。

v0、Screenshot to code、tldraw 都是该形态出色的产品。



v0.dev

Screenshot to code

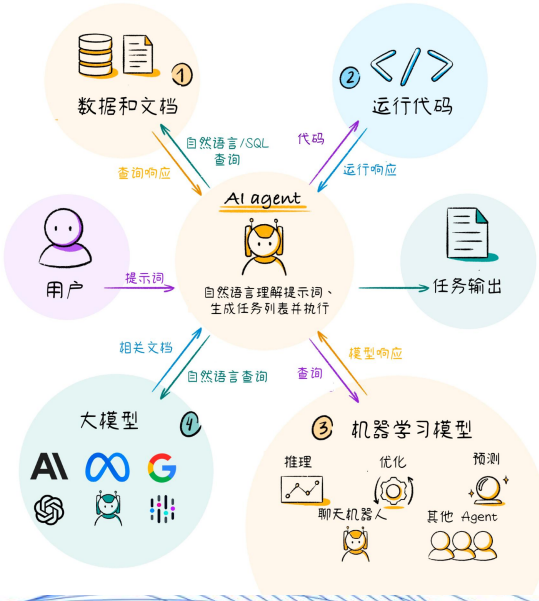


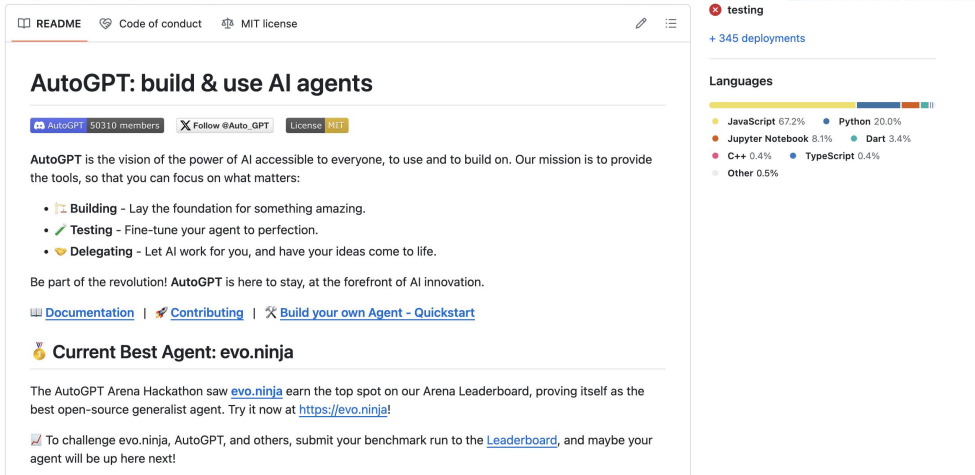
tldraw

LLM Agent（AI Agent）

LLM Agent 是一种基于 LLM 的智能代理，它能够自主学习和执行任务，具有一定的“认知能力和决策能力”。LLM Agent 的出现，标志着 LLM 从传统的模型训练和应用模式，转向以 Agent 为中心的智能化模式。

LLM Agent 打破了传统 LLM 的被动性，使 LLM 能够主动学习和执行任务，从而提高了 LLM 的应用范围和价值；它为 LLM 的智能化发展提供了新的方向，使 LLM 能够更加接近于人类智能。





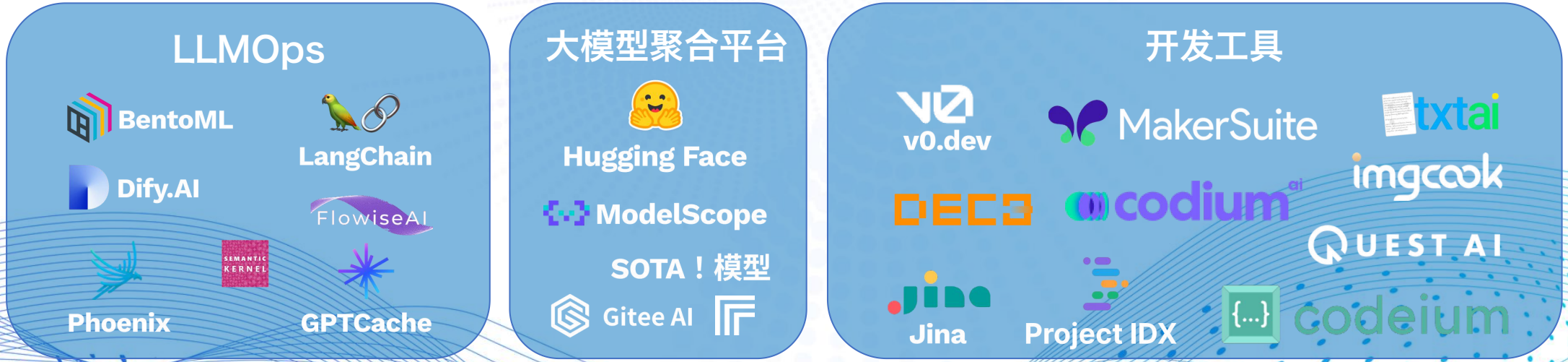
AutoGPT 就是一个典型的 LLM Agent。在给定 AutoGPT 一个自然语言目标后，它会尝试将其分解为多个子任务，并在自动循环中使用互联网和其他工具来实现该目标。它使用的是 OpenAI 的 GPT-4 或 GPT-3.5 API，是首个使用 GPT-4 执行自主任务的应用程序实例。

AutoGPT 最大的特点在于能根据任务指令自主分析和执行，当收到一个需求或任务时，它会开始分析这个问题，并且给出执行目标和具体任务，然后开始执行。

图源：https://blog.abacus.ai/blog/2023/08/31/supercharge-productivity-accomplish-10x-more-with-ai-agents/

LLM 的工具和平台

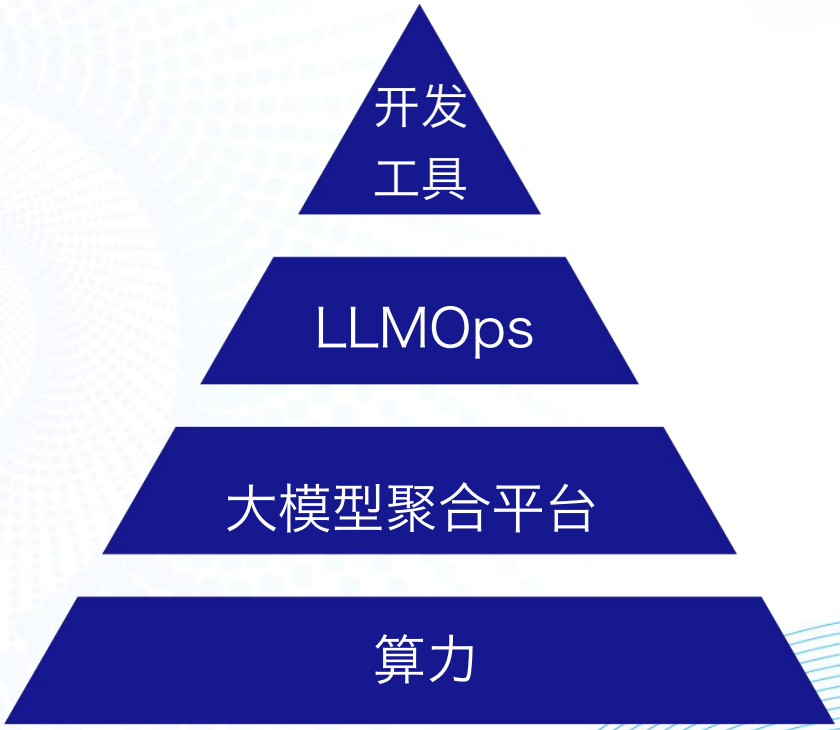
- **LLMOps**：LLMOps 平台专注于提供大模型的部署、运维和优化服务，旨在帮助企业和开发者更高效地管理和使用这些先进的 AI 模型，快速完成从模型到应用的跨越，如 Dify.AI、LangChain 等。
- **大模型聚合平台**： 大模型聚合平台主要用于整合和管理多个大型机器学习模型，在聚合平台之上，衍生出 MaaS（Model-as-a-Service，大模型即服务）的服务模式——通过提供统一的接口和框架，以更高效地部署、运行和优化这些模型，Hugging Face、Replicate 以及 Gitee AI 均为 MaaS 平台。
- **开发工具**：其它开发相关的 LLM 工具，如云原生构建多模态AI应用的工具 Jina，嵌入式数据库 txtai 等。



LLM 的工具、平台和资源

另一个视角来看，在大模型繁荣发展的背后，少不了工具和平台的发力，如 LLMOps 平台、大模型聚合平台以及相关的开发工具，此外还有它们所依赖的最重要的资源——算力。

在这些工具、平台和资源的有力支撑下，大模型才得以一步一个台阶，引领全球开发者步入一个技术新时代。



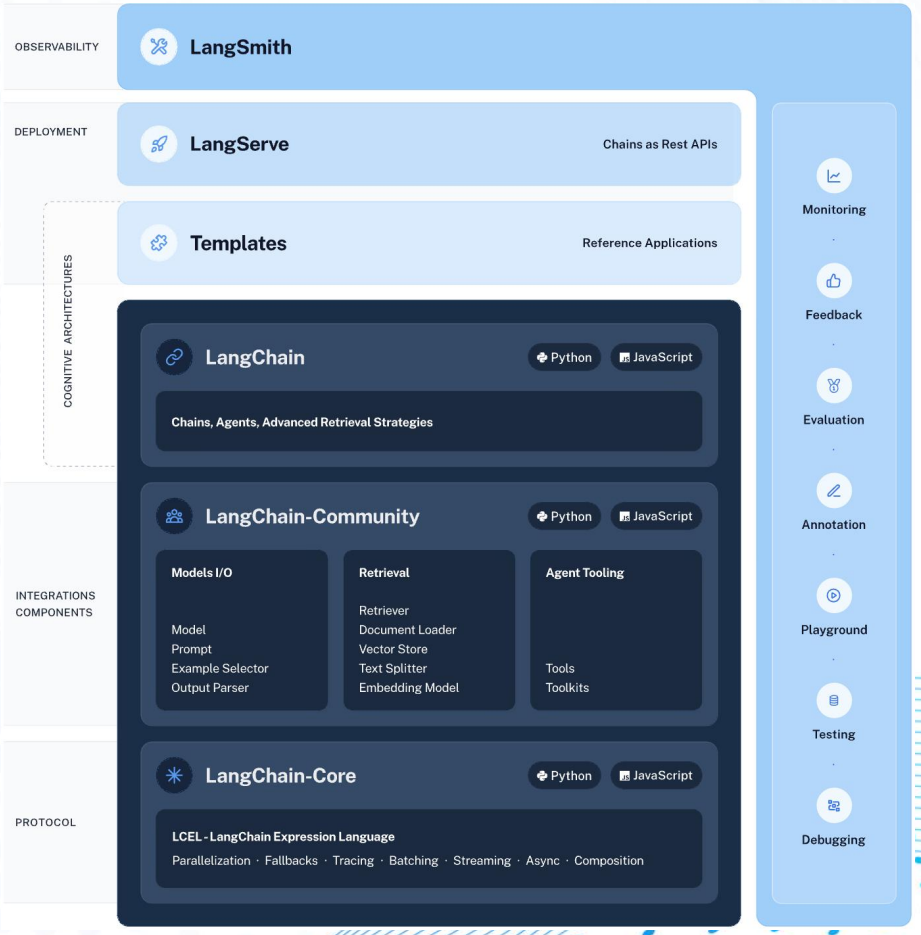
LLM 的工具和平台：LLMOps

开源框架 LangChain

LangChain 是一个帮助开发者使用 LLM 创建应用的开源框架，它可以将 LLM 与外部数据源进行连接，并允许与 LLM 进行交互。

LangChain 于 2022 年 10 月作为开源项目推出，并于 2023 年 4 月注册成立公司，累计获得超过 3000 万美元的投资，估值达到了 2 亿美元。

在 GitHub 上，LangChain 已经获得了超过 7 万个 Star 和超过 2000 名贡献者。

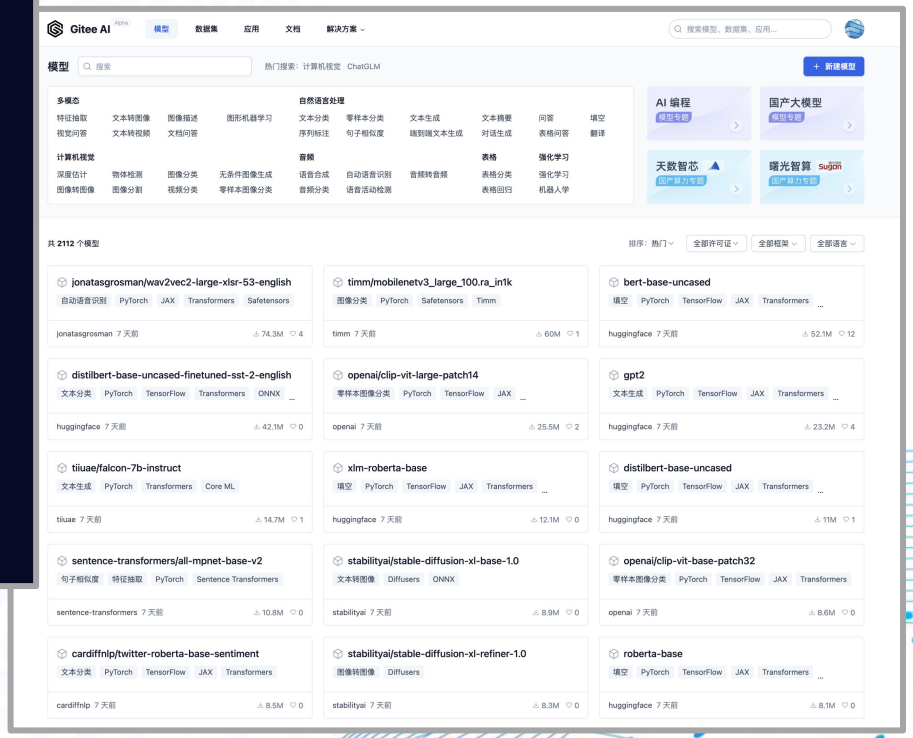


LangChain 架构图

（图源：https://python.langchain.com/docs/get_started/introduction）

LLM 的工具和平台：MaaS 平台

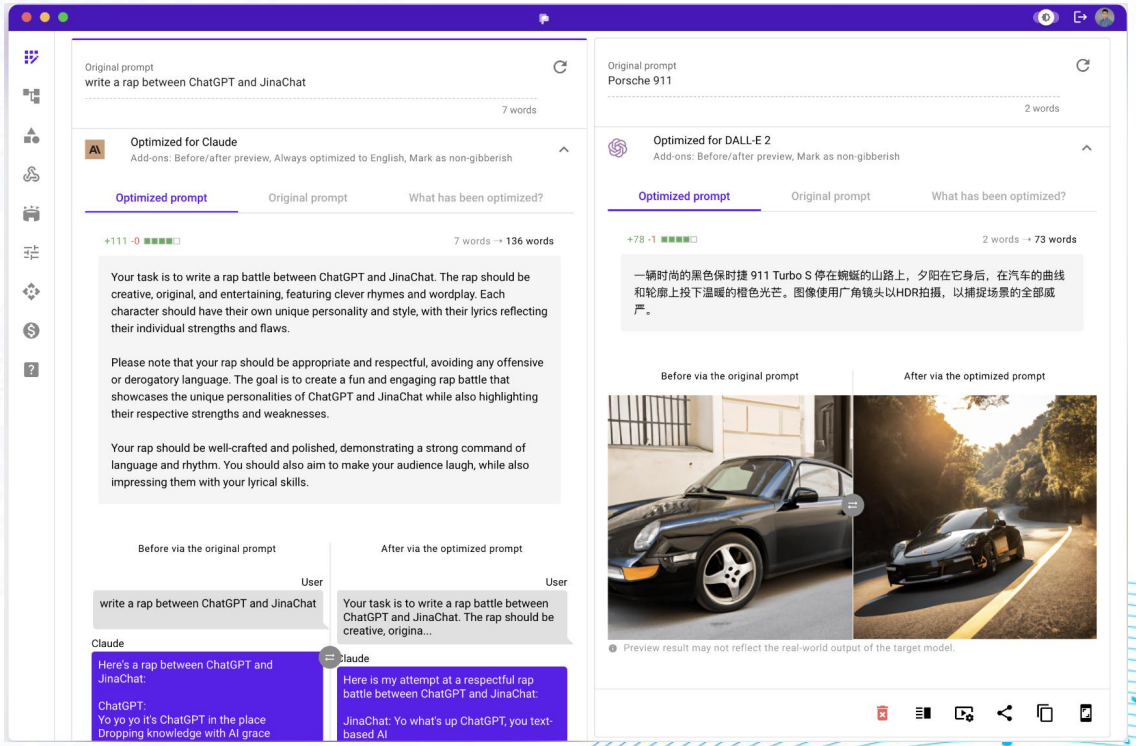
Gitee AI 是开源中国旗下的 MaaS 平台，提供模型、数据集，以及应用托管能力，同时接入了丰富的国产算力平台，为开发者提供了更高效、实惠的微调方案，降低使用门槛，目前已进入内测阶段。



LLM 的工具和平台：开发工具

比较有代表性的 LLM 开发工具有：

- **PromptPerfect**：帮助用户极致优化给大模型的提示词（prompt），使得对大语言模型提问时，可以获得更理想的输出。
- **txtai**：用于语义搜索、LLM 编排和语言模型工作流的一体化嵌入数据库，可以使用 SQL、对象存储、主题建模、图形分析和多模态索引进行矢量搜索。
- **imgcook**：专注以 Sketch、PSD、静态图片等形式的视觉稿作为输入，通过智能化技术一键生成可维护的前端代码，包含视图代码、数据字段绑定、组件代码、部分业务逻辑代码。



PromptPerfect

LLM 世界的基石：算力

LLM 的算力指的是执行这些模型所需的计算资源。这包括用于训练和运行模型的硬件（如 GPU 或 TPU）、内存、存储空间以及处理大量数据的能力。LLM 需要非常强大的算力来处理、理解和生成文本，因为它们涉及到数十亿甚至数万亿个参数的训练和推理。

LLM 的基石是算力，而算力的基石是硬件，硬件的性能直接影响着计算任务的速度、效率和能力。

- **NVIDIA** 是全球领先的 GPU 制造商，提供了强大的图形处理单元，专门用于深度学习和AI计算。
- **华为昇腾系列（HUAWEI Ascend）** AI 处理器和基础软件构建 Atlas 人工智能计算解决方案，打造面向“端、边、云”的全场景 AI 基础设施方案，覆盖深度学习领域推理和训练全流程。
- **AMD** 被外界视为打破 NVIDIA 垄断 AI 算力市场的多一种选择，其基于第三代 CDNA 架构，为生成式 AI 大语言模型设计的 MI300X 内存高达 192GB，集成了高达 1530 亿个晶体管，为历代产品之最。
- **昆仑芯** 科技团队自研，面向通用AI计算的芯片核心架构昆仑芯 XPU 从AI落地的实际需求出发，按照复杂前沿的人工智能场景需求开展迭代，致力为开发者提供通用、易用、高性能的算力来源。
- **海光** DCU 系列产品以 GPGPU 架构为基础，兼容通用的“类 CUDA”环境以及国际主流商业计算软件和人工智能软件，可广泛应用于大数据处理、人工智能、商业计算等应用领域。
- **天数智芯** 通用 GPU 高端芯片及超级算力系统提供商。拥有云边协同、训推组合的完整通用算力系统全方案，其系统架构、指令集、核心算子、软件栈均为自主研发，可独立发展演进。

LLM 世界的基石：算力

算力也是全国乃至世界范围内 LLM 相关企业遇到的最大难题：

➢ 资源少

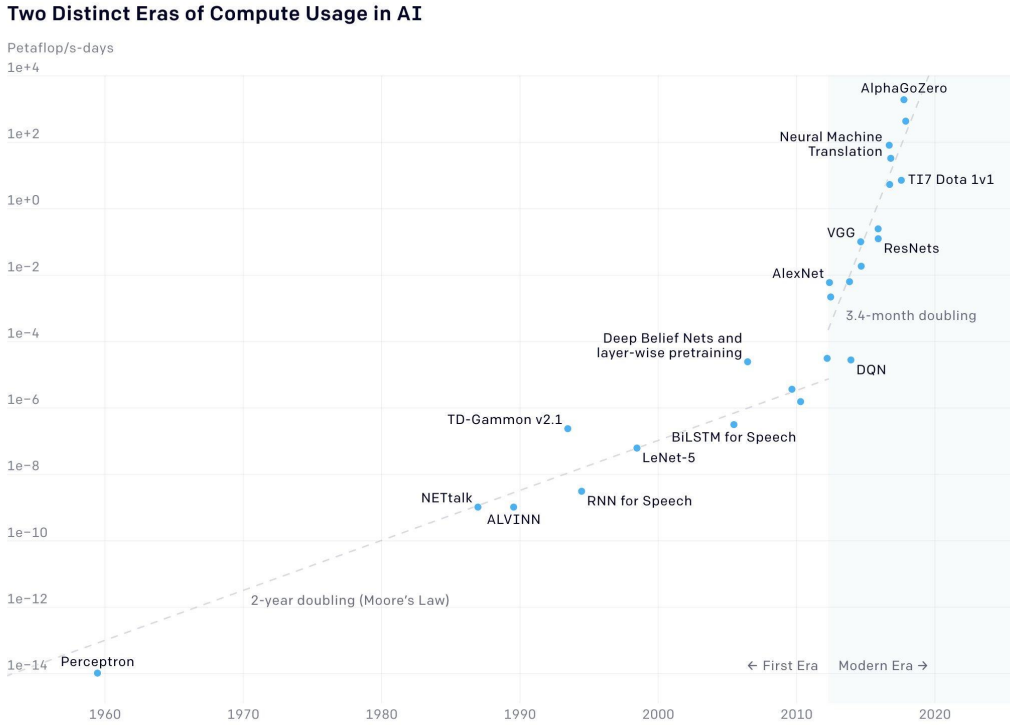
随着国内大模型数量激增，AI 算力需求从 2022 年开始持续上涨，国内市场出现一卡难求的情况。根据 IDC 预计，到 2026 年 AI 推理的负载比例将进一步提升至62.2%，特别是预训练大模型几乎成为 AI 开发的标准范式。同时，这一需求也导致了 NVIDIA A100 GPU 的价格在几个月内暴涨超过 50%，而且大量断货。

➢ 成本高

根据研究测算，单次 GPT- 3 模型（175B）训练，在规模 300B token 下成本约为 35000 卡·天（A100），也就是相当于 35000 块 A100 GPU 跑 1 天能完成单次训练，或者 2500 块 A100 GPU 跑 2 周。以每张卡 10 万人民币的价格计算，单次训练成本就将达到 25-35 亿人民币。

➢ 客观限制

10月17日，美国商务部工业和安全局(BIS)公布新的先进计算芯片、半导体制造设备出口管制规则，限制中国购买和制造高端芯片的能力，受管制的包括但不限于 NVIDIA A100、H100、A800、H800、L40、L40S 以及集成这些高性能计算的 DGX/HGX 系统，并将中国 GPU 企业及其子公司列入了实体清单。



据 OpenAI 测算，自 2012 年以来，人工智能模型训练算力需求每3~4个月就翻一番，每年训练 AI 模型所需算力增长幅度高达 10 倍
（图源：https://openai.com/research/ai-and-compute）

查看完整报告：
https://talk.gitee.com/report/china-open-source-2023-annual-report.pdf

Thank You

公众号

视频号

关注我们，开源开发者圈一网打尽

OSCHINA oschina.net

gitee gitee.com