第19期--极光月全食

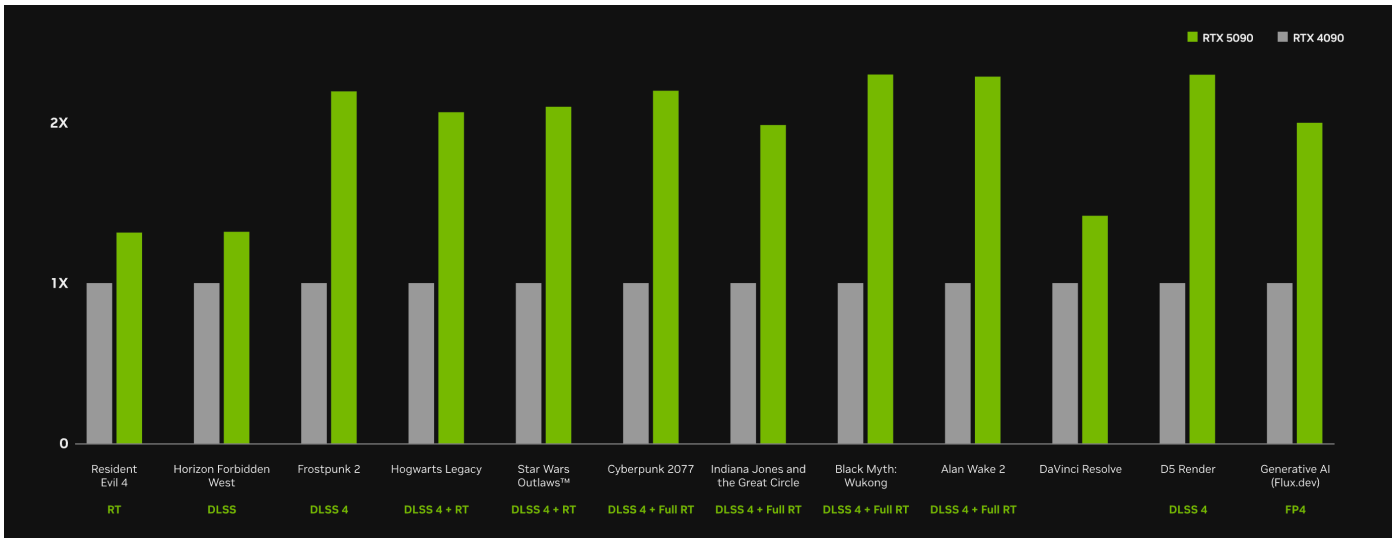
一、刊首图

一个美国摄影师前往阿拉斯加州拍摄极光，他无意中发现，这段时间还有月全食，于是成功拍到了极光月全食。月全食的时候，月球、地球、太阳成一条直线，月球落在地球的阴影里面，照不到直接的太阳光，而是被地球大气层反射的太阳光照亮。地球反射的是太阳光的红光，所以月全食呈现红色，又称“血月”。这张绿色极光中的“血月”照片，非常难得。

二、时事新闻

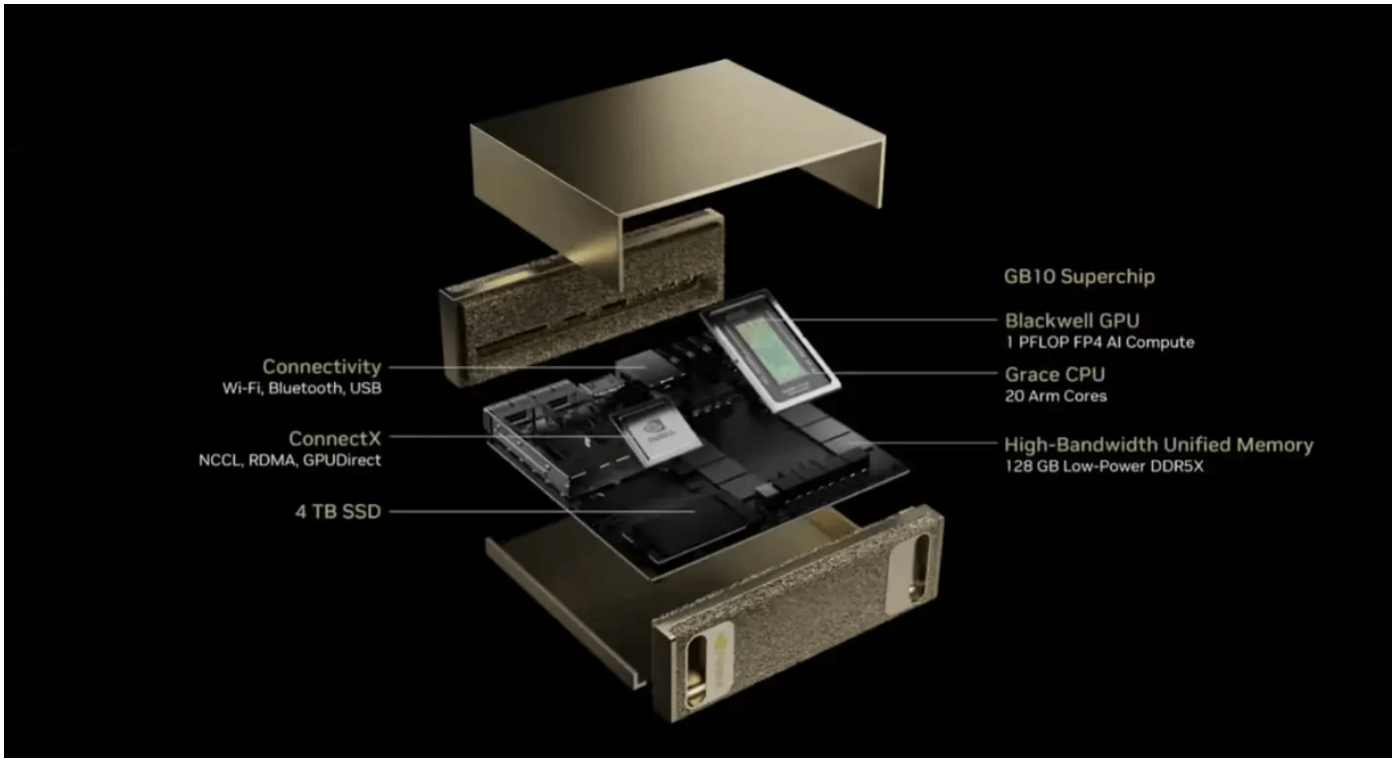
1、GeForce RTX 5090显卡

NVIDIA RTX 5090搭载了最新的NVIDIA Blackwell架构，这是NVIDIA为提升图形计算能力而推出的新一代GPU架构。相比前代的Ada Lovelace架构，Blackwell在性能、功耗效率、以及AI计算方面都有显著提升。



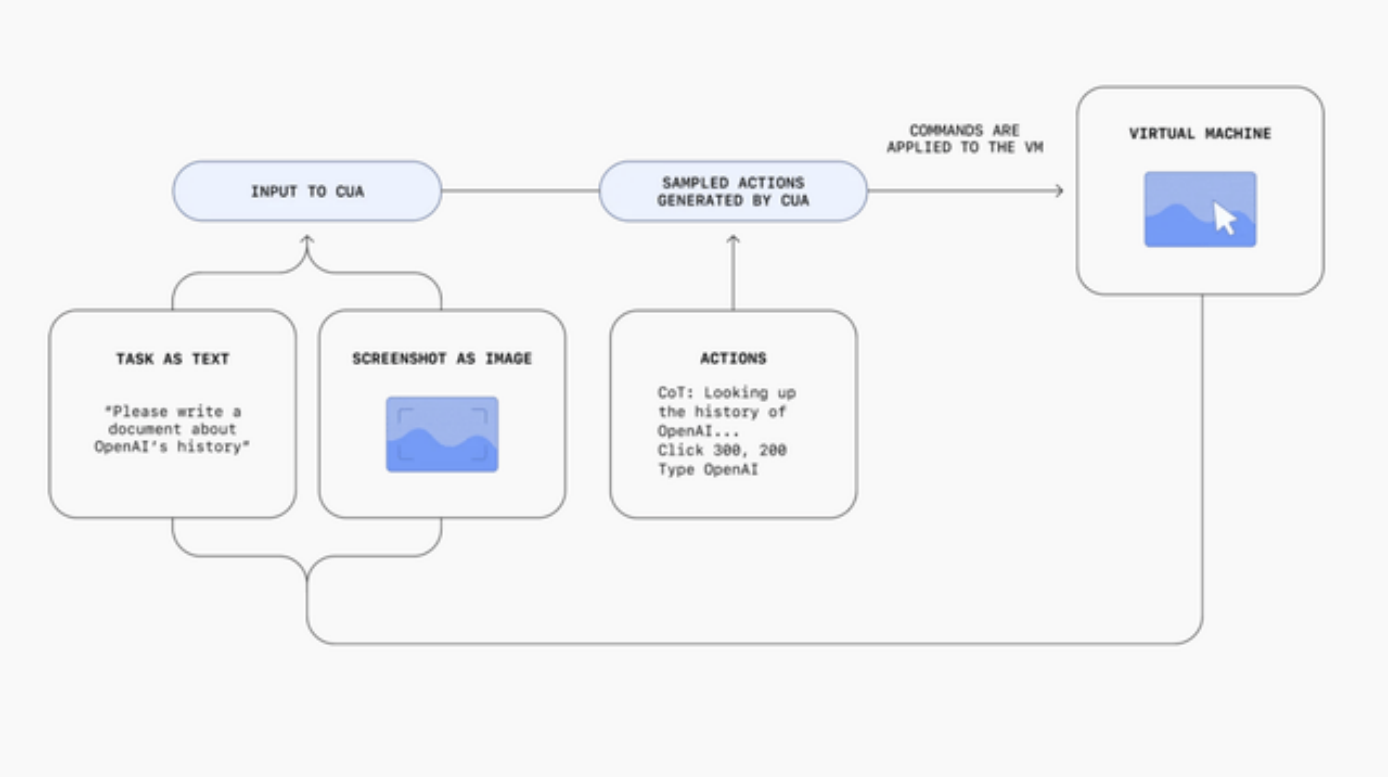
2、[Project Digits本地超算]

英伟达宣布将于5月推出名为“Project Digits”的个人AI超算，起价3000美元。该设备大小类似Mac Mini，配备GB10 Grace Blackwell超级芯片，能够运行最多2000亿参数的AI模型，支持最多128GB统一内存和4TB NVMe存储。通过连接两个单元，能够处理最多4050亿参数的模型。Project Digits基于Linux操作系统，预装NVIDIA AI软件栈，适用于开发者和AI研究人员，提供本地开发、测试和部署功能。



3、[OpenAI发布Operator智能体]

OpenAI于1月23日发布了“Operator”智能体，能通过上网为用户完成任务，如订餐、购物和预订票务。它将结合GPT-4的视觉和推理能力，能通过截图和浏览器交互，自动执行任务，遇到隐私信息时暂停等待用户确认。目前已给美国地区的ChatGPT Pro用户推送了该功能，但该智能体仍是研究预览版，功能有限。

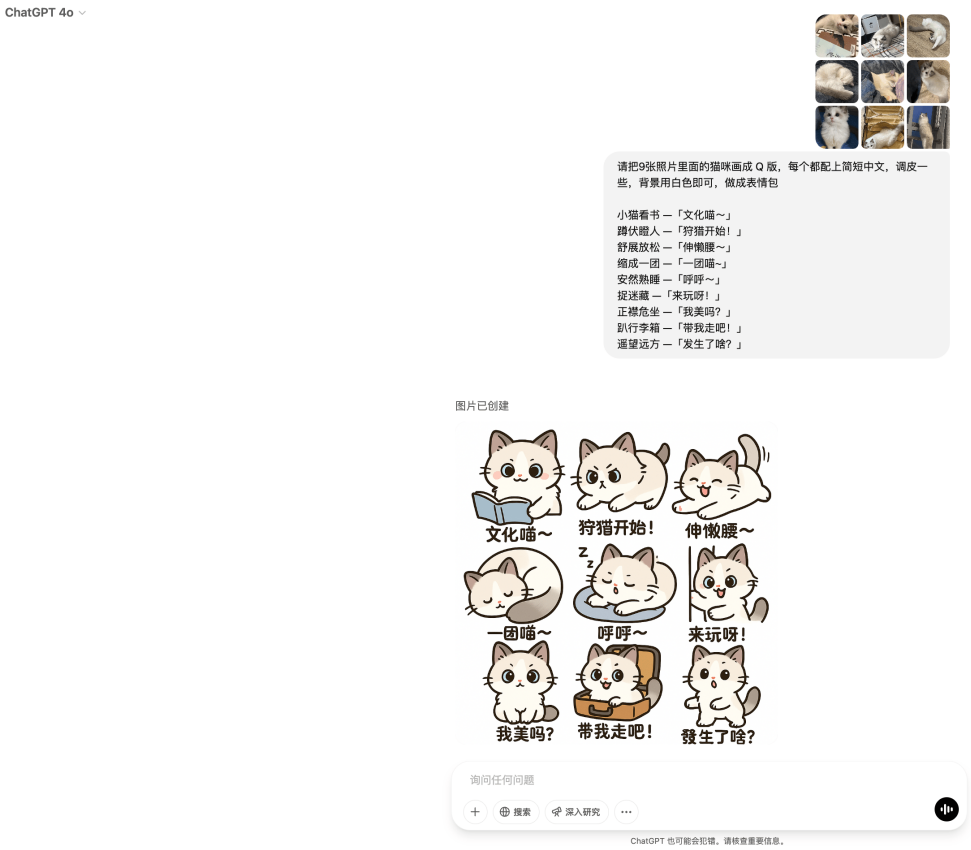


4、Android15的终端程序

Android15 将有一个原生的终端程序，提供一个基于 Debian 的 Linux 发行版供用户使用。这个功能的底层是虚拟机机制，它将大大方便程序员，将安卓手机当作 Linux 桌面电脑使用。

5、GPT-4o原生图像生成

2025.3.26，GPT-4o支持了原生出图。与DALL-E 3不同，此次OpenAI的全新图像生成器基于其原生多模态GPT-4o模型，能够同时理解图像和文本，可以非常好地遵循提示词指令，轻松创作出虚实结合的场景，就像在现实中一样。有网友发现，用它生成表情包效果极佳。



6、谷歌发布Gemini 2.5模型

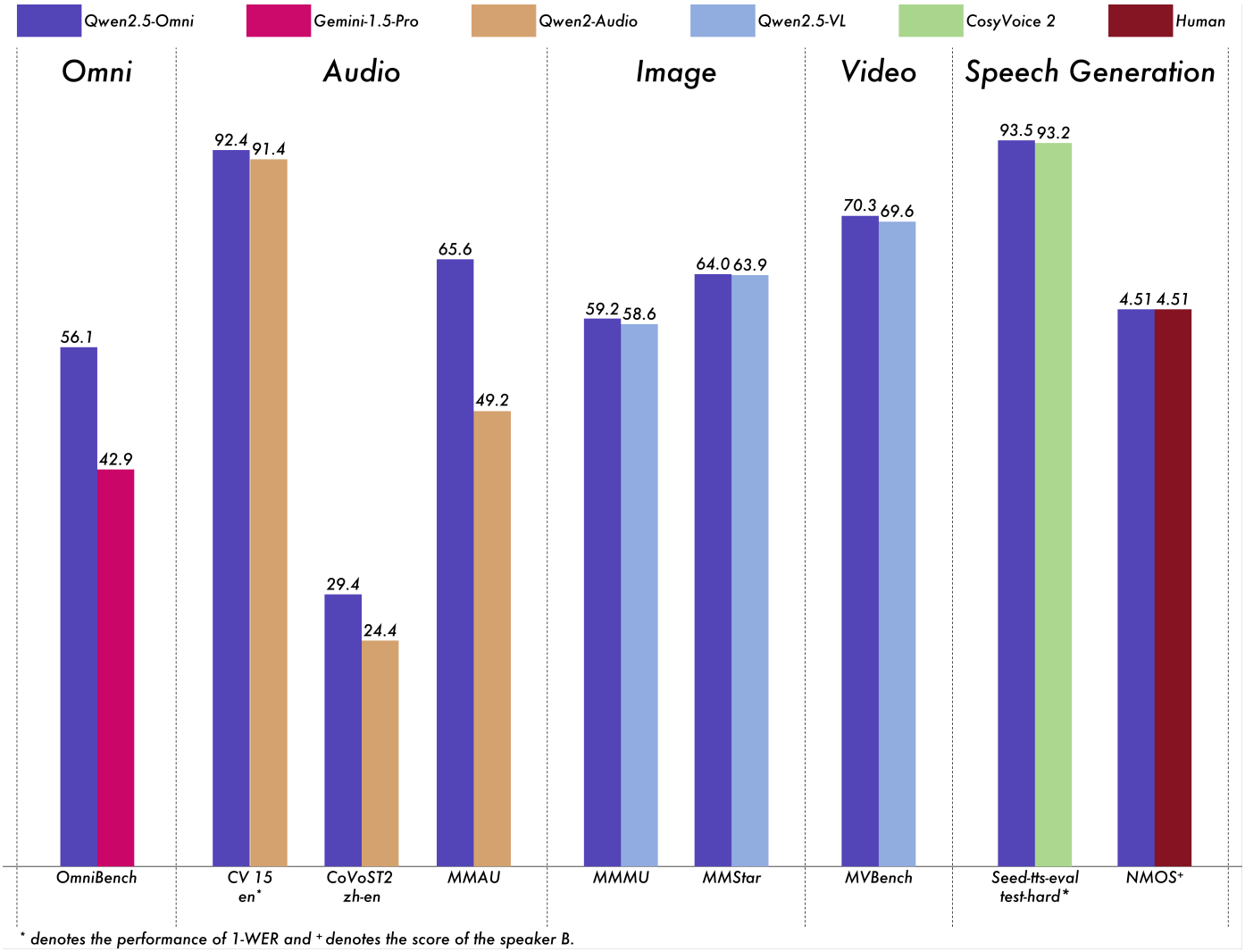
Gemini 2.5 Pro是一个‘思考’模型，能够在回应前先进行思考推理，从而提升性能，并改善准确性。它在多个基准测试中达到了SOTA水平，并且以显著的优势在LMArena上排名第一。

Benchmark		Gemini 2.5 Pro Experimental (03-25)	OpenAI o3-mini High	OpenAI GPT-4.5	Claude 3.7 Sonnet 64k Extended Thinking	Grok 3 Beta Extended Thinking	DeepSeek R1
Reasoning & knowledge Humanity's Last Exam (no tools)		**18.8%**	14.0%*	6.4%	8.9%	—	8.6%*
Science GPQA diamond	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	**84.8%**	84.6%	—
Mathematics AIME 2025	single attempt (pass@1)	**86.7%**	86.5%	—	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	**93.3%**	—
Mathematics AIME 2024	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	—	—	—	80.0%	**93.3%**	—
Code generation LiveCodeBench v5	single attempt (pass@1)	70.4%	**74.1%**	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	**79.4%**	—
Code editing Aider Polyglot		**74.0% / 68.6%** whole / diff	60.4% diff	44.9% diff	64.9% diff	—	56.9% diff
Agentic coding SWE-bench verified		63.8%	49.3%	38.0%	**70.3%**	—	49.2%
Factuality SimpleQA		52.9%	13.8%	**62.5%**	—	43.6%	30.1%
Visual reasoning MMMU	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	—	no MM support	—	—	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		69.4%	no MM support	—	—	—	no MM support
Long context MRCR	128k (average)	**94.5%**	61.4%	64.0%	—	—	—
	1M (pointwise)	83.1%	—	—			
Multilingual performance Global MMLU (Lite)		89.8%	—	—	—	—	—

7、发布Qwen2.5-Omni

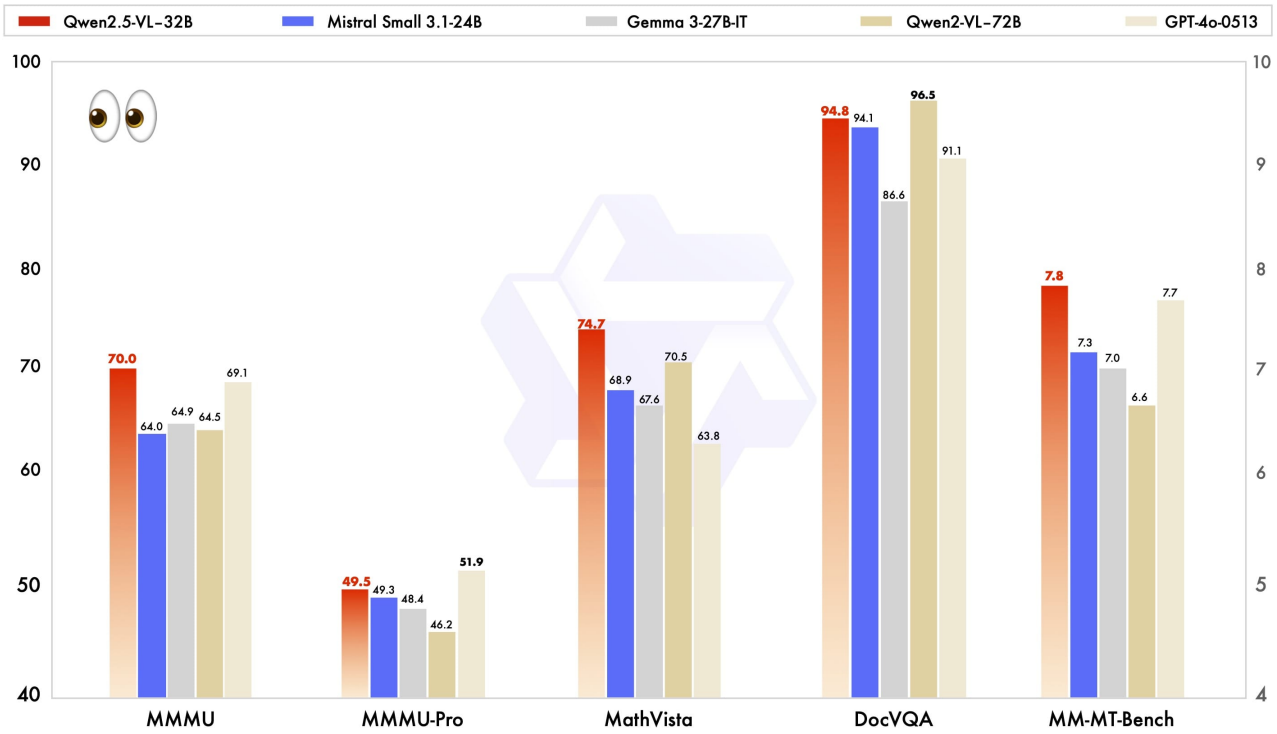
阿里发布并开源首个端到端全模态大模型——通义千问Qwen2.5-Omni-7B，仅靠一个一体式模型，就能搞定文本、音频、图像、视频全模态，并实时生成文本和自然语音。

在多模态任务OmniBench评测中，Qwen2.5-Omni表现刷新记录拿下新SOTA，远超谷歌Gemini-1.5-Pro等同类模型。在单模态的语音识别、翻译、音频理解、图像推理、视频理解、语音生成任务中，Qwen2.5-Omni的全维度表现也都优于类似大小的单模态模型以及闭源模型。在seed-tts-eval语音生成基准中，Qwen2.5-Omni展现出与人类水平相当的语音合成能力。



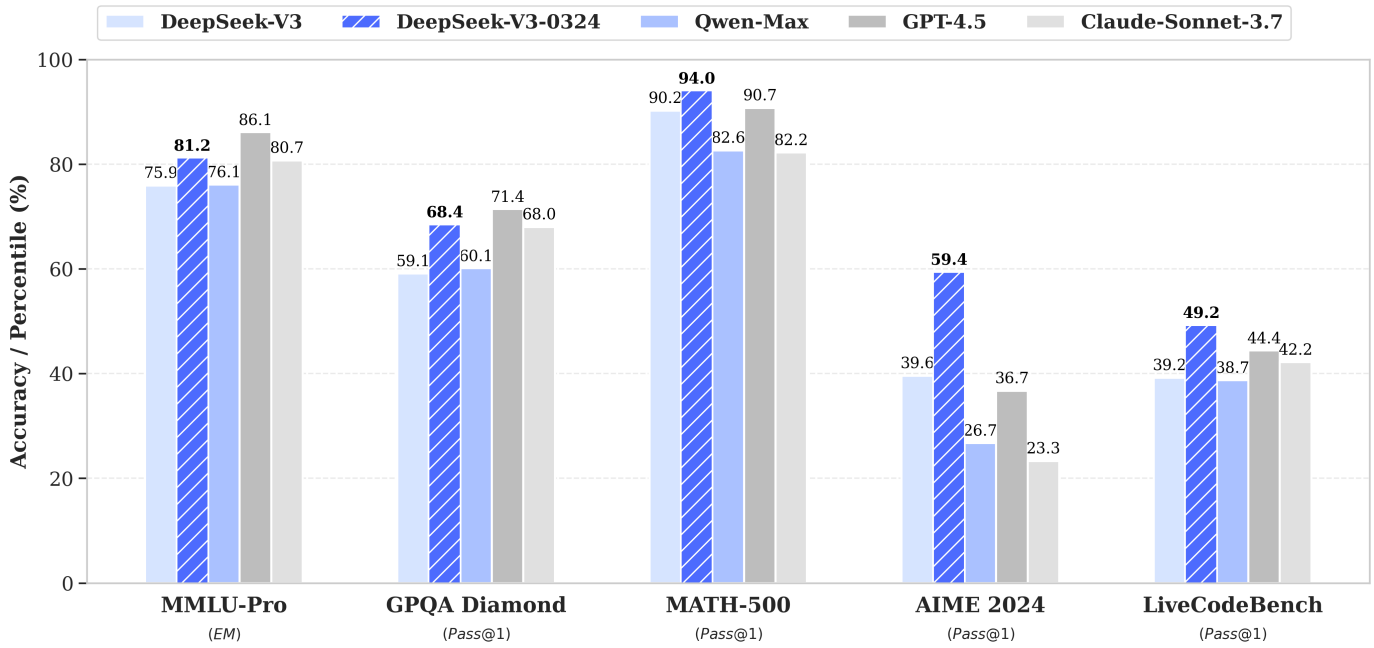
8、发布Qwen2.5-VL-32B-Instruct

发布全新Qwen2.5-VL-32B-Instruct模型，针对视觉任务进行了强化学习优化，显著提升了人类偏好对齐和数学推理能力。模型在图像解析、内容识别和视觉逻辑推理方面表现优异，性能超过部分更大规模的模型，如Qwen2-VL-72B-Instruct。



9、[发布DeepSeek-V3-0324]

DeepSeek发布了新版本模型 DeepSeek-V3-0324，该版本借鉴了DeepSeek-R1中的强化学习技术，并在多个任务上进行了深度优化，在推理能力、数学与代码方面实现了显著提升，整体性能超越GPT-4.5。



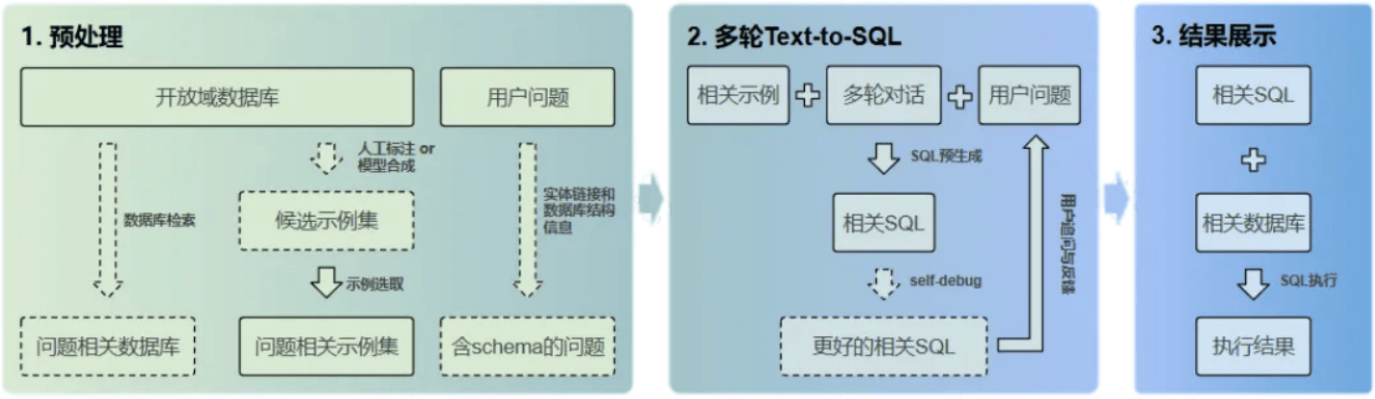
三、技术文章

1、[Qwen2.5-LLM：扩展大型语言模型的边界]（中文）

阿里通义千问官方发布的文章，其中有关于Qwen2.5大模型介绍、与其他模型的推理效果对比。

2、[哈工大SCIR发布珠算-SQL]（中文）

珠算-SQL是由SCIR实验室开发的Text-to-SQL系统，旨在将自然语言查询自动转换为SQL查询，支持多轮对话、自动领域迁移和多数据库检索，具备用户反馈纠正机制，提供高效精准的数据库交互体验，本文介绍了其技术路线。



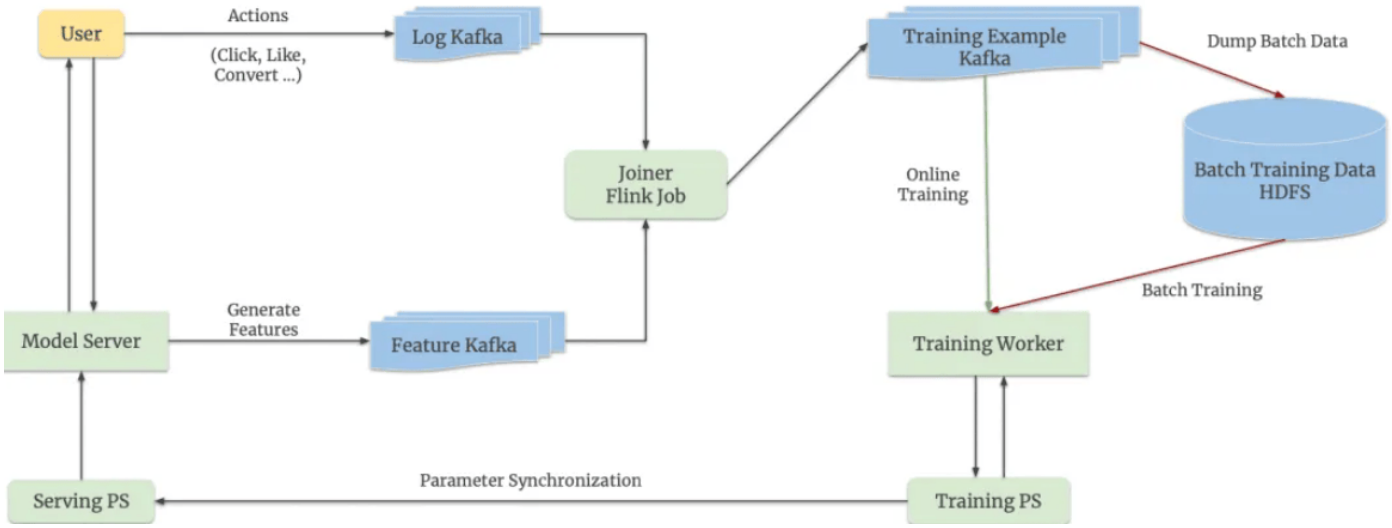
3、[如何用 Claude Code 反编译代码](#)（英文）

作者演示了一个例子，使用 Anthropic 发布的 Claude Code，将 Webpack 编译出来的文件反编译，还原成源代码。

四、开源组件

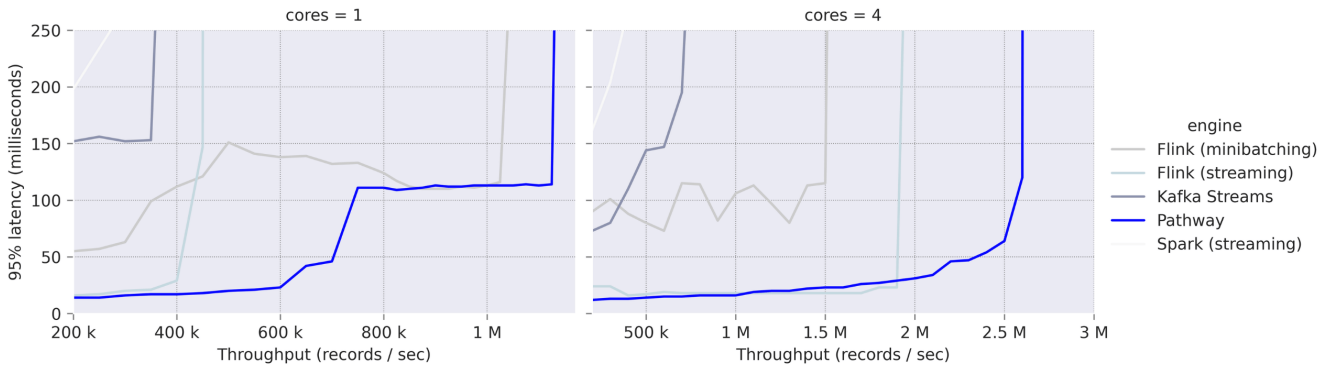
1、[monolith](#)

Monolith 是一个基于 TensorFlow 构建的大规模推荐模型深度学习框架，支持批量和实时训练及服务。



2、[pathway](#)

Pathway 是一个 Python ETL 框架，用于流处理、实时分析、LLM 管道和 RAG。尽管 Pathway 是用 Python 编写的，但它是由 Rust 引擎运行，从而实现多线程、多处理和分布式计算。



The graphs report the latency in milliseconds achieved by each benchmarked platform for a given throughput value.
Pathway outperforms state-of-the-art solutions for common data streaming tasks across both low and high throughput values.
Lower is better.

3、Qwen-Agent

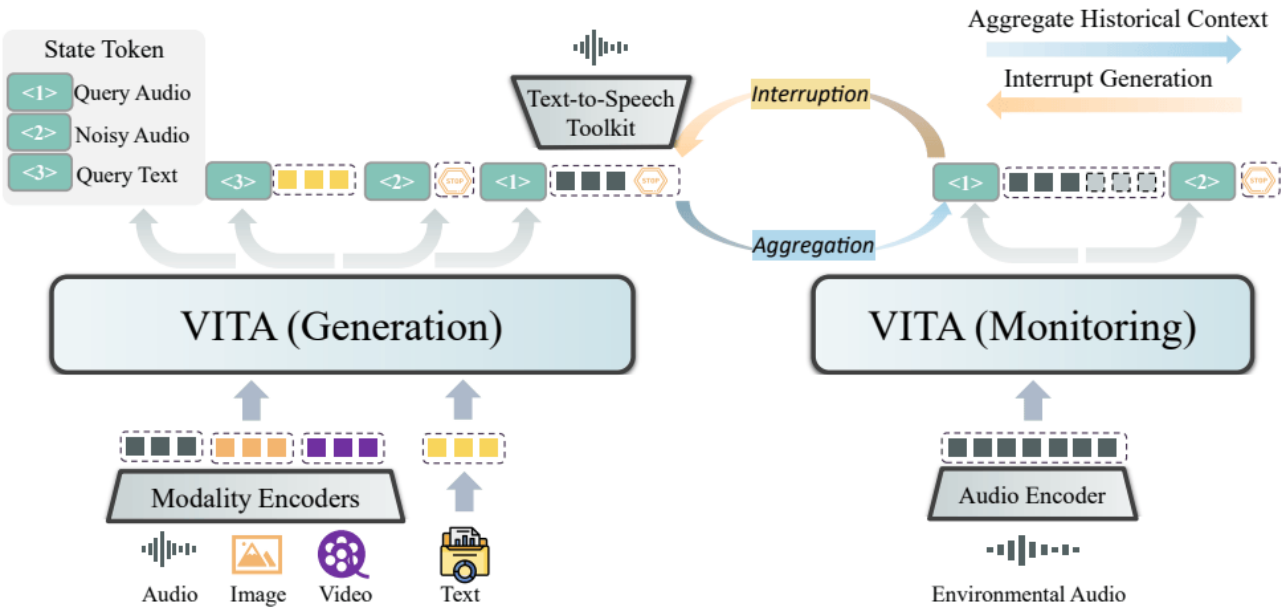
Qwen-Agent是一个开发框架，可用来开发Agent应用，充分利用基于Qwen模型的指令遵循、工具使用、规划、记忆能力。本项目也提供了浏览器助手、代码解释器、自定义助手等示例应用。

4、agibot-world

AgiBot World数据集诞生于智元自建的大规模数据采集工厂与应用实验基地，一方面为机器人大规模数据训练提供场地，另一方面真实复刻了家居、餐饮、工业、商超和办公五大核心场景，全面覆盖了机器人在生产、生活中的典型应用需求。

5、VITA

VITA是腾讯优图实验室推出的首个开源多模态大语言模型，支持同时处理视频、图像、文本和音频。它具备中英文双语理解与生成能力，提供无唤醒交互和音频打断等自然交互功能。论文地址



6、outlines

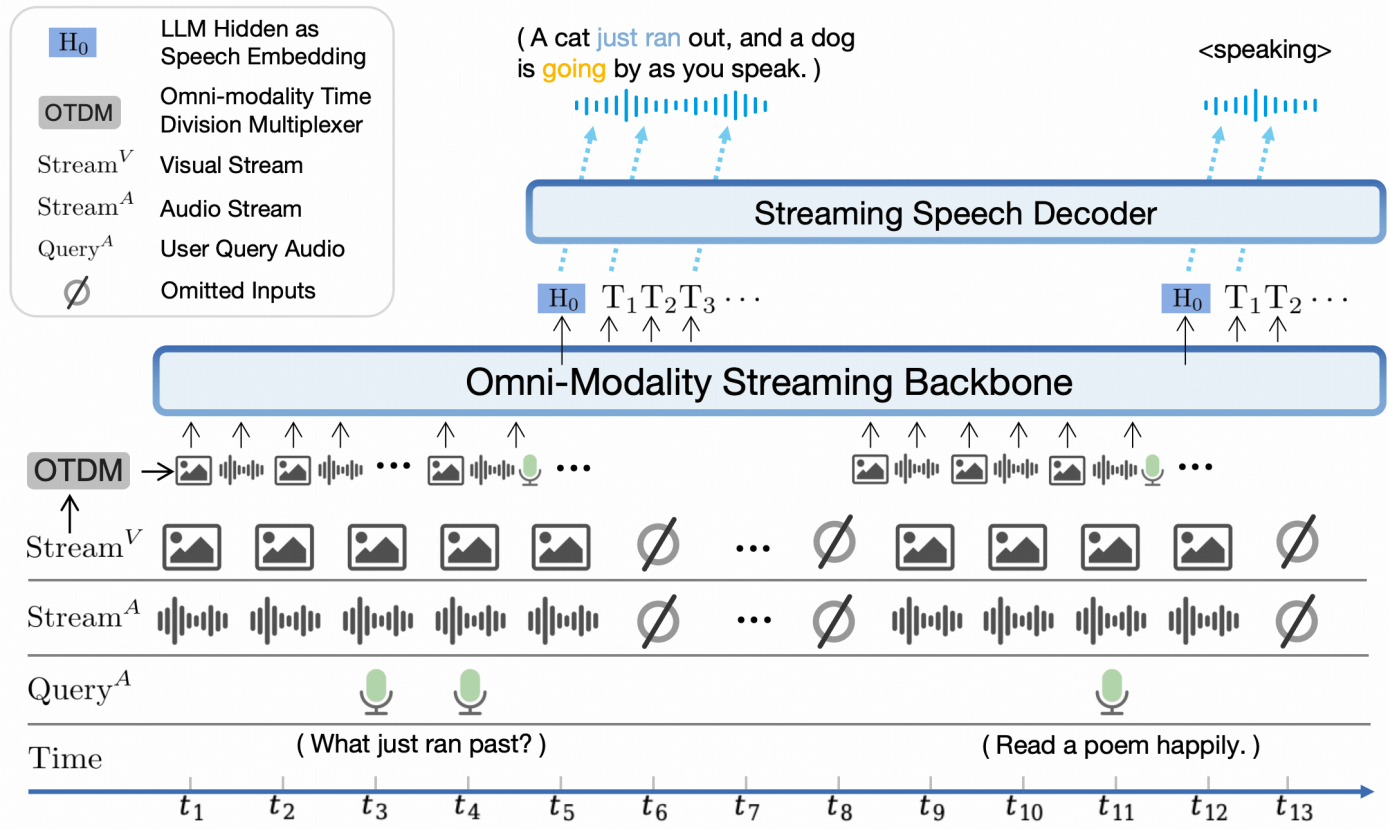
Outlines 是一个帮助用户简单稳定地使用 LLM 的 Python 库，支持基于正则表达式、JSON 和语法规则实现结构化输出。

7、agent-service-toolkit

一个基于LangGraph、FastAPI和Streamlit构建的完整工具集，旨在帮助开发者快速构建和运行AI代理服务。

8、MiniCPM-o

MiniCPM-o 是从 MiniCPM-V 升级的最新端侧多模态大模型系列。该系列模型可以以端到端方式，接受图像、视频、文本、音频作为输入，并生成高质量文本和语音输出。



9、RealtimeSTT

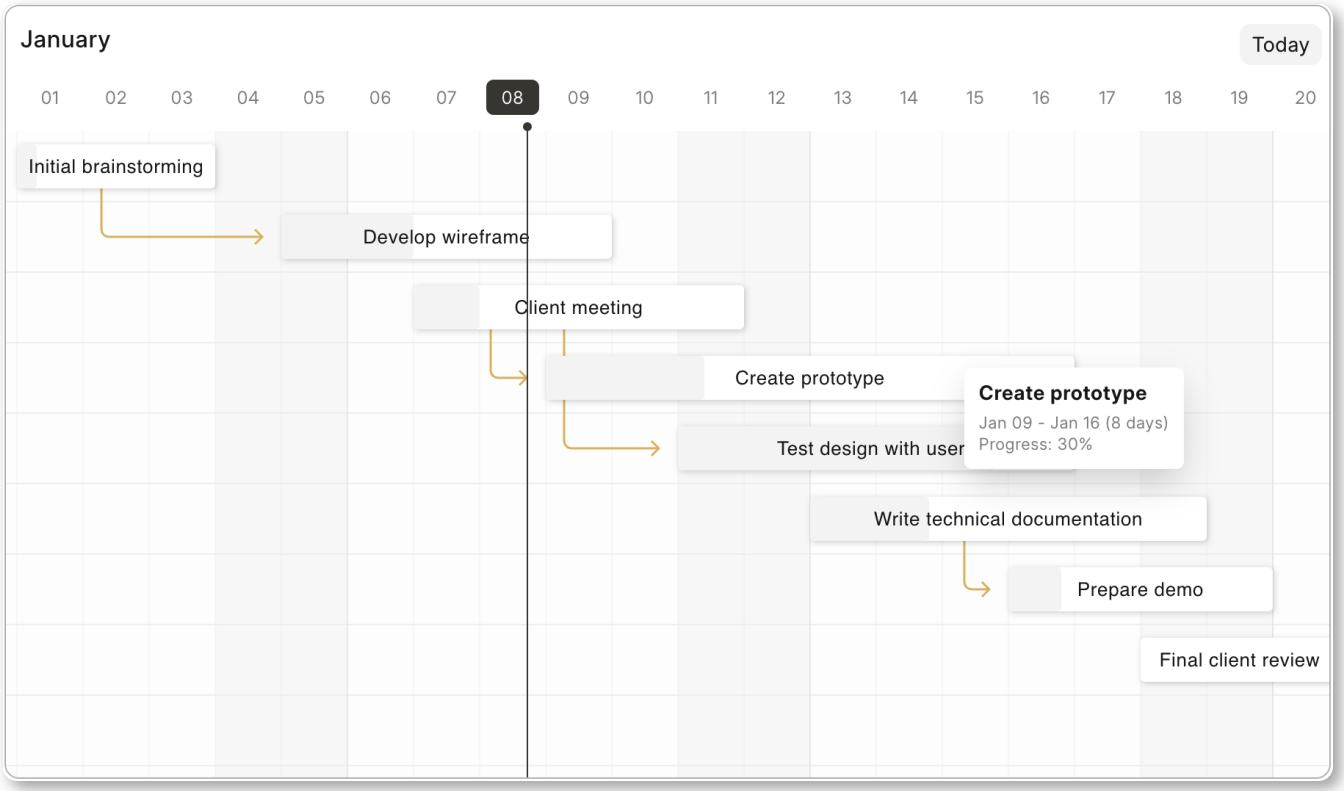
RealtimeSTT是一个强大、高效、低延迟的语音转文本库，具有先进的语音活动检测、唤醒词激活和即时转录功能。

10、sharp

一个基于 libvips 的高性能 Node.js 图像处理库，支持对 JPEG、PNG、WebP、GIF 和 SVG 等格式的图像进行调整大小、格式转换、裁剪和旋转等操作。

11、Frappe Gantt

Frappe Gantt是一款开源的JavaScript 甘特图库，具有简洁的界面和丰富的交互功能。



12、SemHash

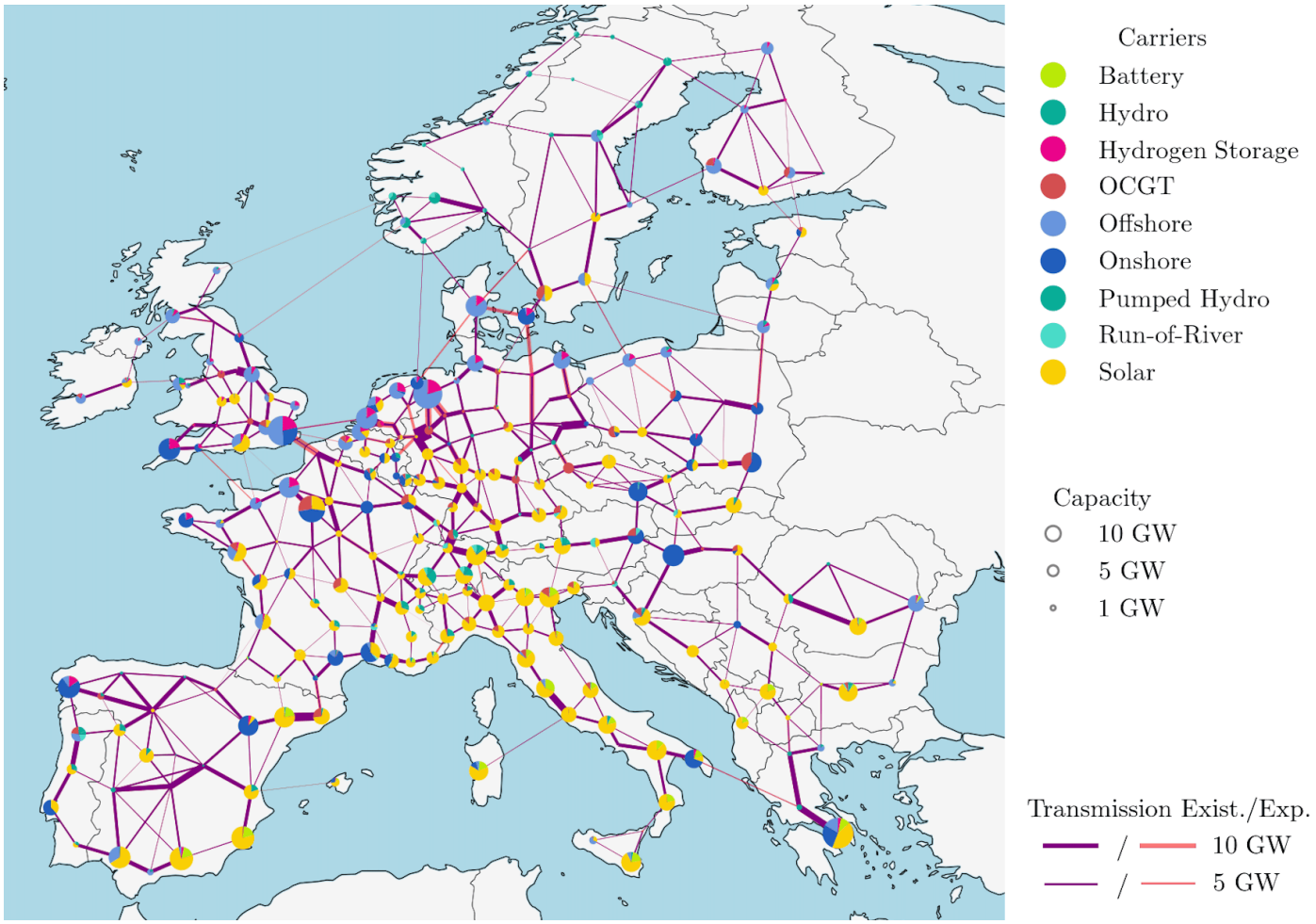
SemHash是一个轻量且灵活的工具，用于通过语义相似性来去重数据集。它结合了Model2Vec的快速嵌入生成和Vicinity的高效ANN相似性搜索，支持单数据集去重（例如，清理训练集）和多数据集去重（例如，确保测试集和训练集之间没有重叠）。

13、chonkie

一个专为 RAG 应用设计的轻量级文本分块库，它简单易用、速度快，能够按固定大小分割文本，支持多种分词器、向量模型和灵活的分块策略，适用于长文本处理、构建 RAG 应用等场景。

14、PyPSA

一个用于电力系统分析的 Python 库，专注于电力和多能源系统的建模与优化。它基于 Pandas、NumPy、GLPK、Cbc 等库，能够高效计算最优潮流优化（OPF）、线性和非线性电力流，并支持模拟各种电力和能源系统组件的功能。

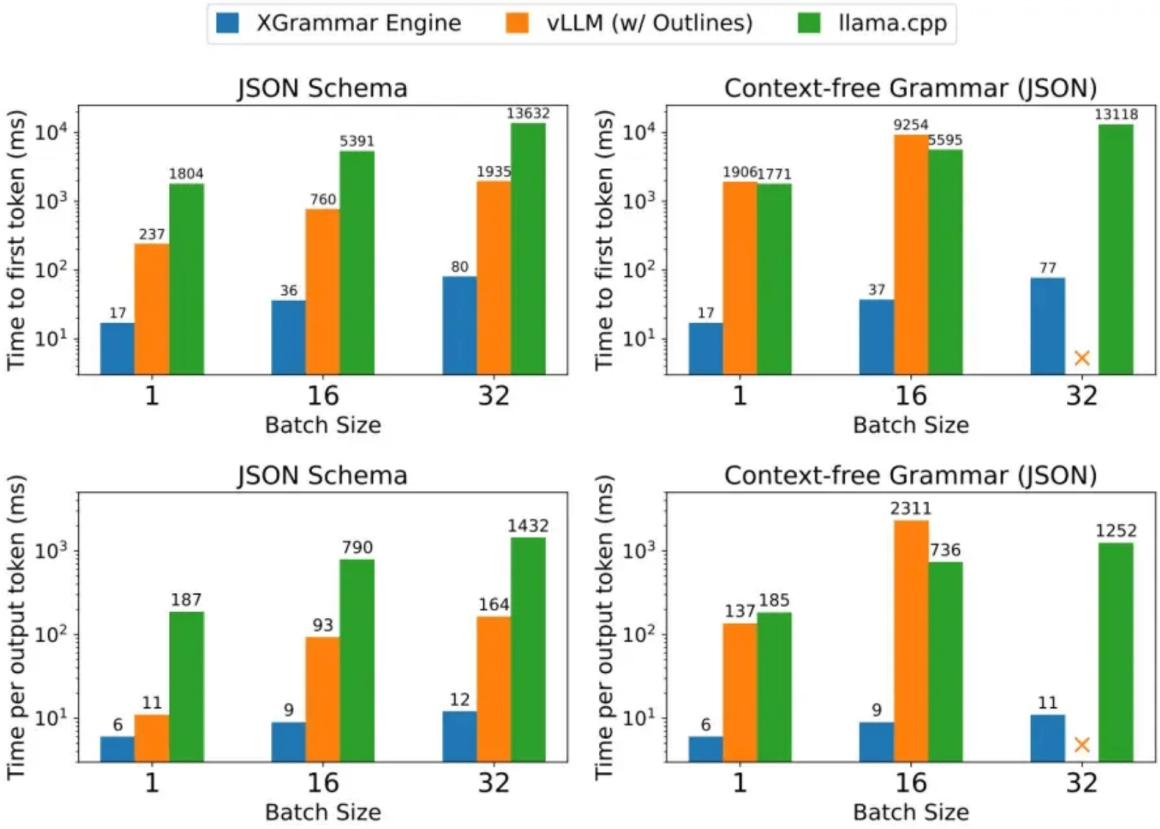


15、DeepSeek-R1

DeepSeek发布DeepSeek-R1模型，并同步开源模型权重。据官方介绍，它在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能比肩OpenAI o1正式版。技术报告

16、XGrammar

XGrammar 结合 LLM 推理引擎，它能够在端到端低延迟 LLM 服务中实现近乎零额外开销的结构化生成。



17、instructor

该项目是用于处理大语言模型结构化输出的 Python 库。它基于 Pydantic 实现了数据验证和类型注释，能够将 LLM 的结果转换为结构化数据，支持多种大语言模型服务，以及自动重试、流式响应等功能。

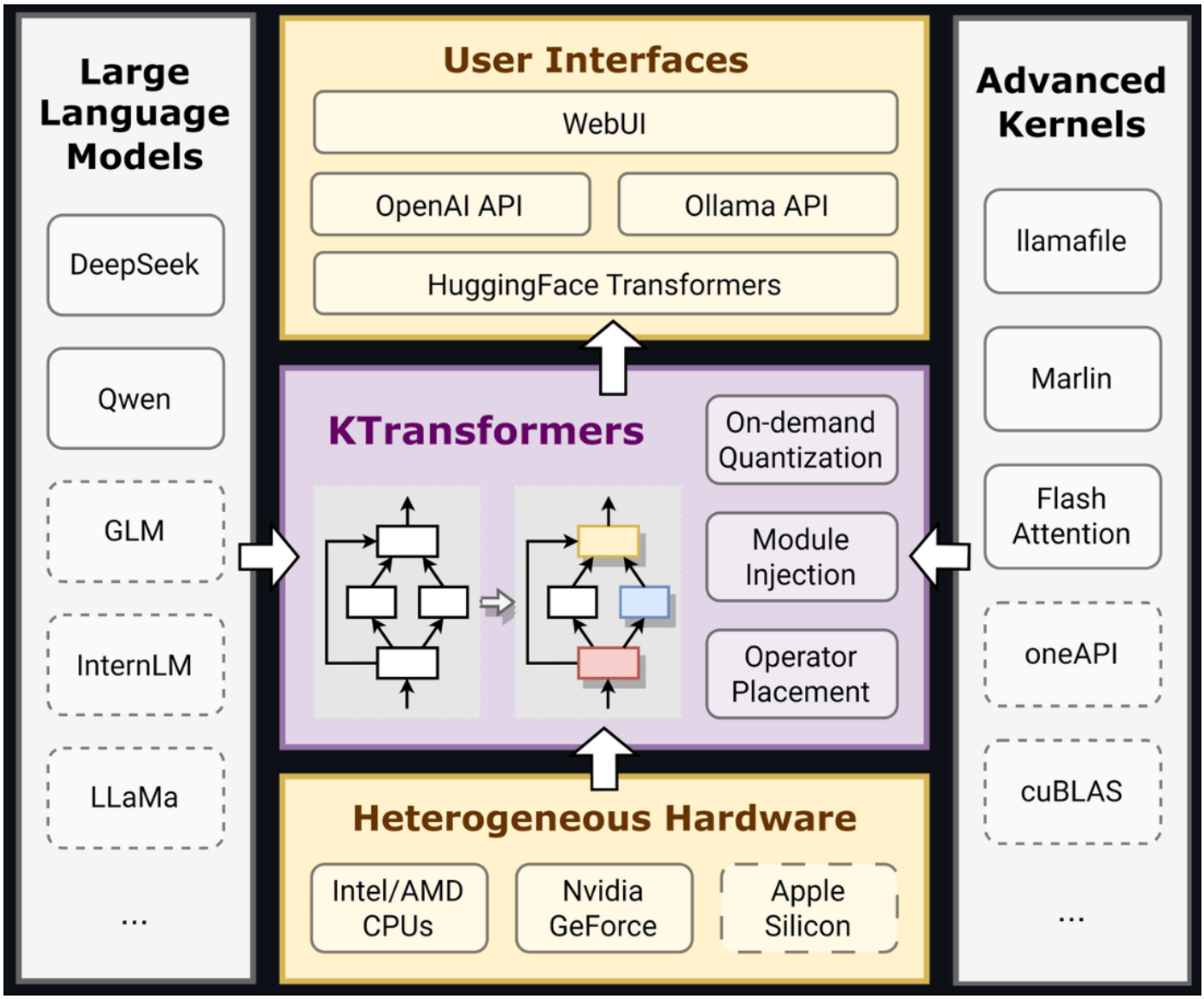
18、Unsloth

Unsloth可以比HuggingFace快2-5倍的微调Llama 3.3、Mistral、Phi-4、Qwen 2.5和Gemma等大语言模型，同时内存消耗减少80%。

Unsloth supports	Free Notebooks	Performance	Memory use
Llama 3.2 (3B)	▶️ Start for free	2x faster	70% less
Phi-4 (14B)	▶️ Start for free	2x faster	70% less
Llama 3.2 Vision (11B)	▶️ Start for free	2x faster	50% less
Llama 3.1 (8B)	▶️ Start for free	2x faster	70% less
Gemma 2 (9B)	▶️ Start for free	2x faster	70% less
Qwen 2.5 (7B)	▶️ Start for free	2x faster	70% less
Mistral v0.3 (7B)	▶️ Start for free	2.2x faster	75% less
Ollama	▶️ Start for free	1.9x faster	60% less
ORPO	▶️ Start for free	1.9x faster	50% less
DPO Zephyr	▶️ Start for free	1.9x faster	50% less

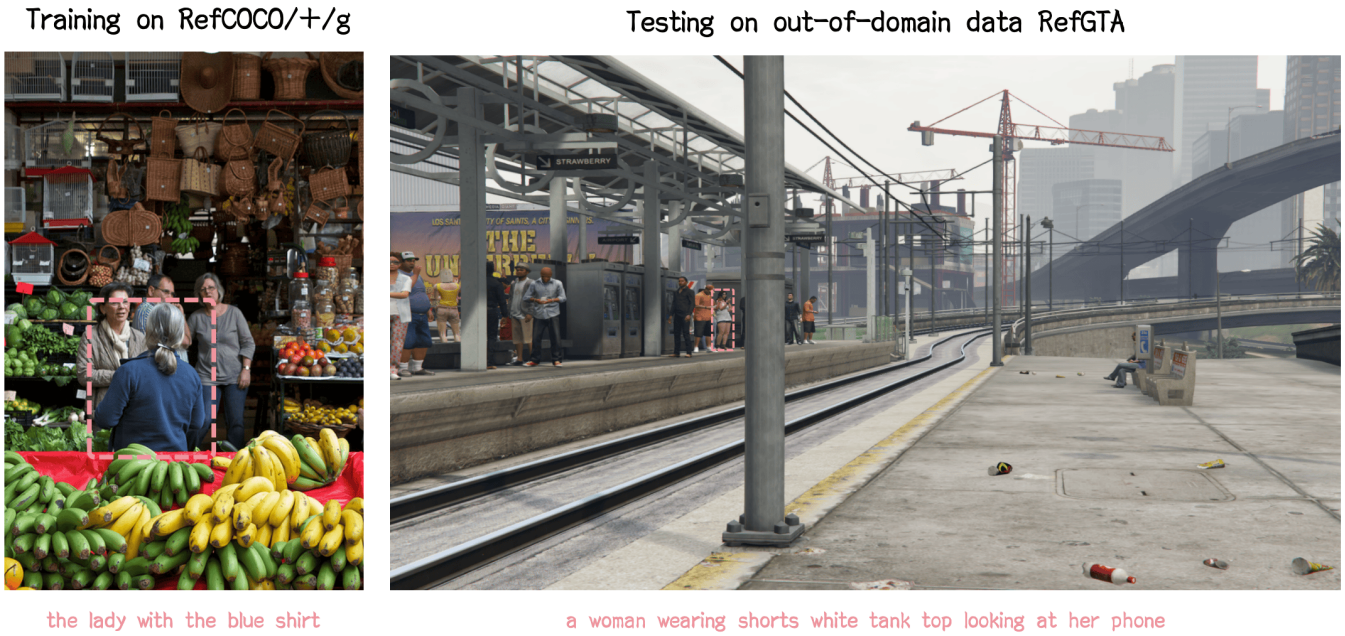
19、KTransformers

KTransformers是一个由清华大学MADSys和Approaching.AI开发的开源框架，旨在优化大模型的本地推理体验，特别是支持DeepSeek-R1等MoE大型语言模型的高效运行。它的性能较llama.cpp大幅提升，尤其在Prefill阶段，速度提升高达27.79倍。



20、VLM-R1

VLM-R1是用强化学习提升视觉理解的大型视觉语言模型。在视觉指代表达理解任务中，R1模型在域外数据上表现稳定，优于SFT模型。

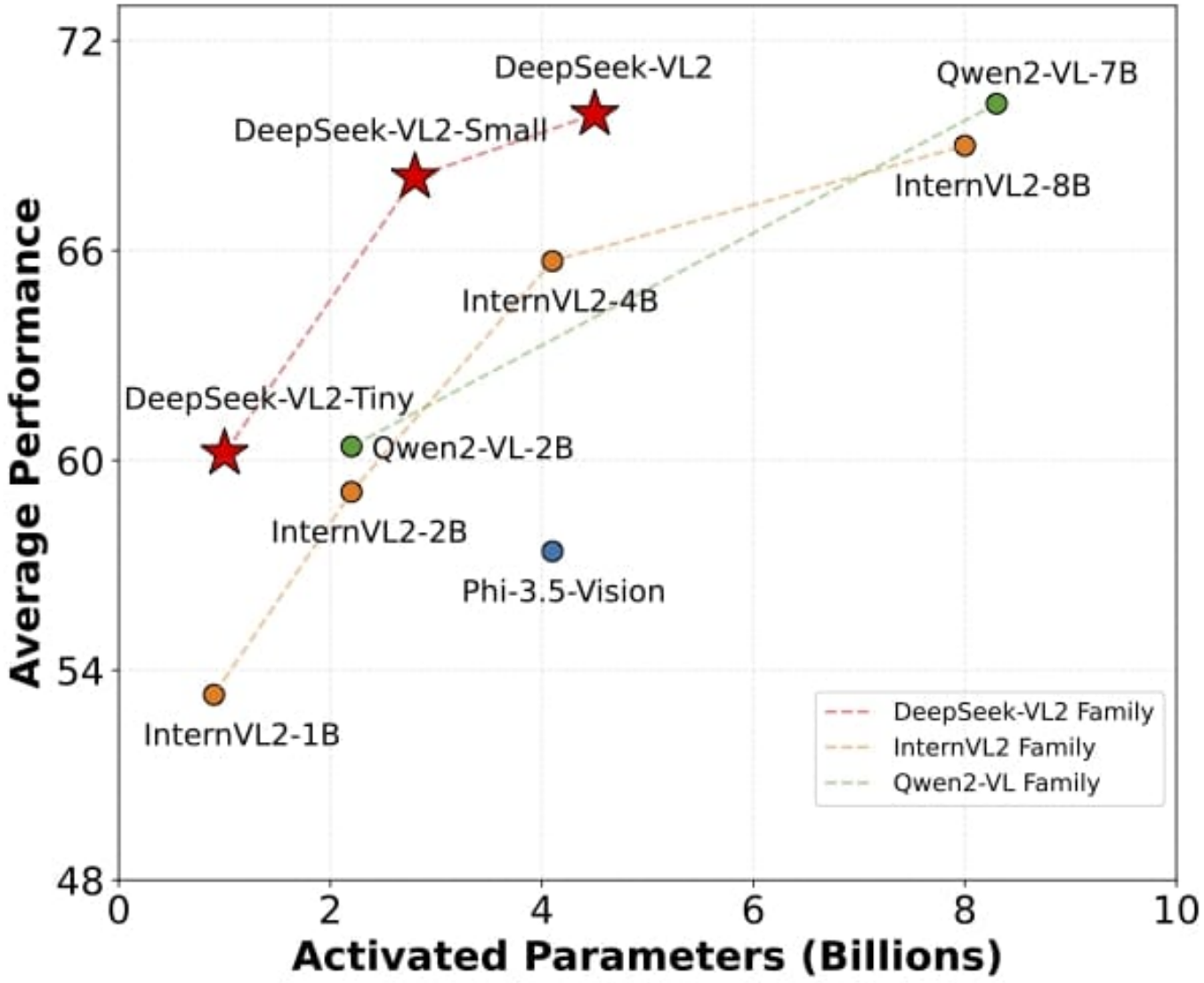


21、champ

一个开源的四足机器人开发框架，可用于构建四足机器人和开发控制算法。它提供轻量级的头文件库和丰富的 ROS 工具包，支持完全自主导航、Gazebo 仿真环境以及多种硬件平台，为开发者提供了完整的四足机器人控制框架和开发工具。

22、DeepSeek-VL2

DeepSeek-VL2是一个大型混合专家（MoE）视觉语言模型系列，显著改进了其前身DeepSeek-VL。DeepSeek-VL2在各种任务中表现出色，包括但不限于视觉问答、光学字符识别、文档/表格/图表理解和视觉定位。

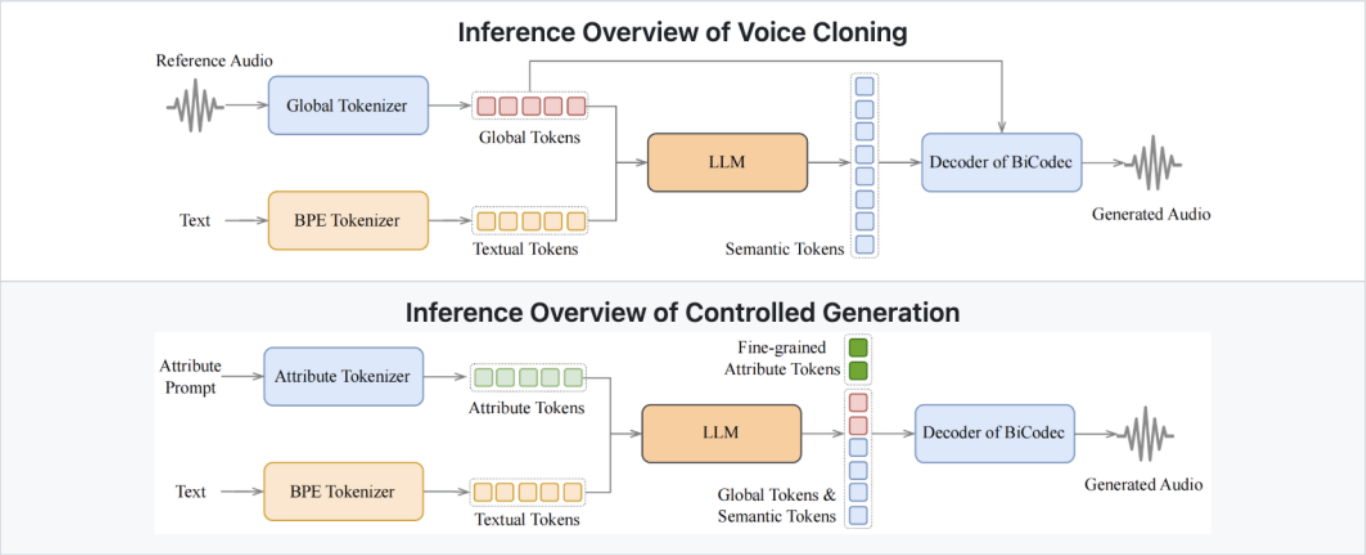


23、open-infra-index

DeepSeek 开源的 open-infra-index 项目为 AGI 研究和开发提供了重要资源。该团队公开了5个经过生产环境验证的存储库，并配备完善的文档与部署支持，展现了他们在 AGI 领域的关键进展。

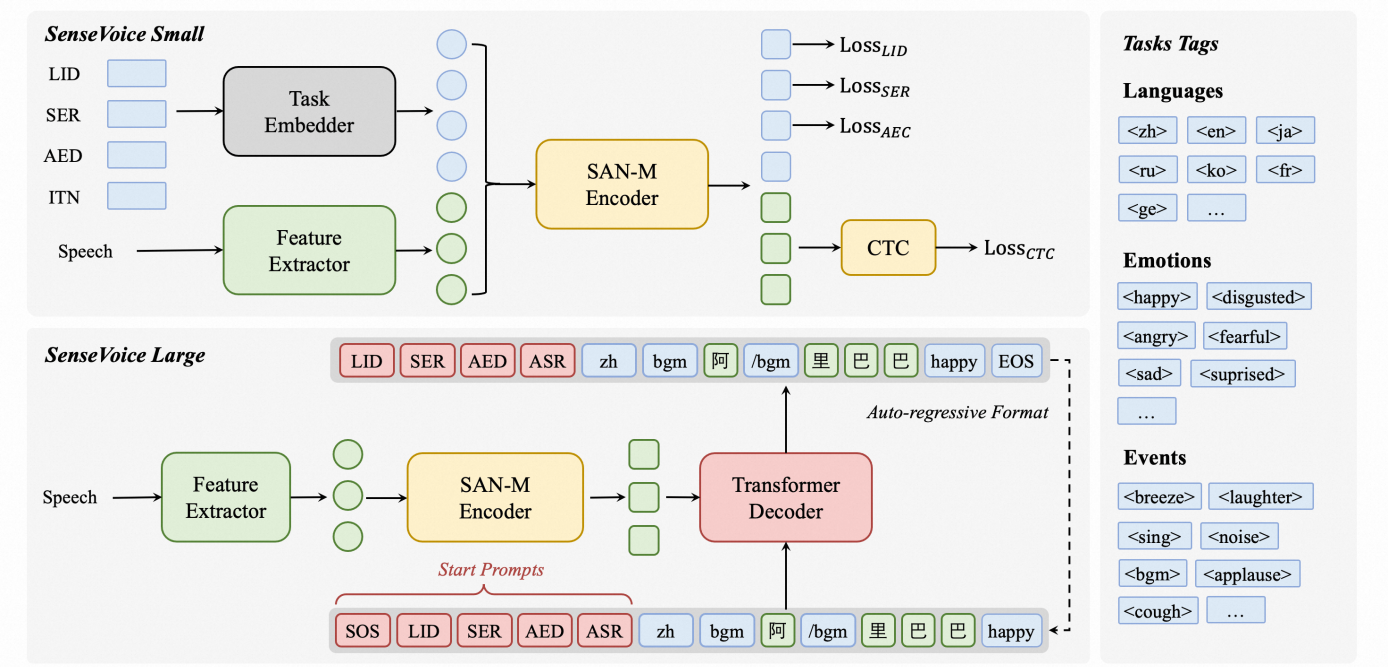
24、Spark-TTS

Spark-TTS 是一款高质量语音合成（TTS）系统，它不仅支持零样本语音克隆，还能进行细粒度语音控制，包括语速、音调、语气等多项参数调节，同时具备跨语言生成能力，让 AI 语音变得更加灵活、多样化。



25、SenseVoice

SenseVoice 是具有音频理解能力的音频基础模型，包括语音识别（ASR）、语种识别（LID）、语音情感识别（SER）和声学事件分类（AEC）或声学事件检测（AED）。



五、工具软件

1、EmojiClick（免费）

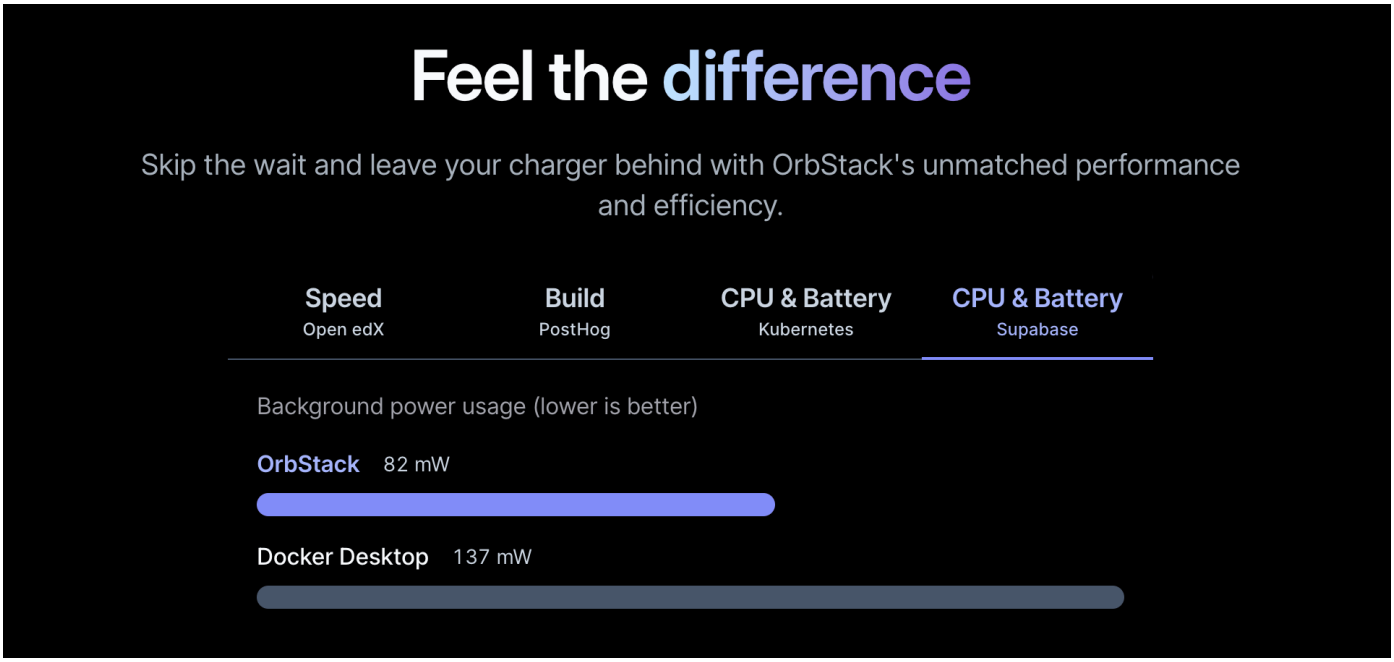
使用自然语言搜索 Emoji 符号。

2、AutoMouser（开源）

一个 Chrome 浏览器插件，将鼠标操作通过 AI 转为 Selenium Python 脚本。

3、OrbStack（免费）

OrbStack是一款专为macOS设计的快速、轻量级且易于操作的Docker容器工具，官方声称它比DockerDesktop速度更快。

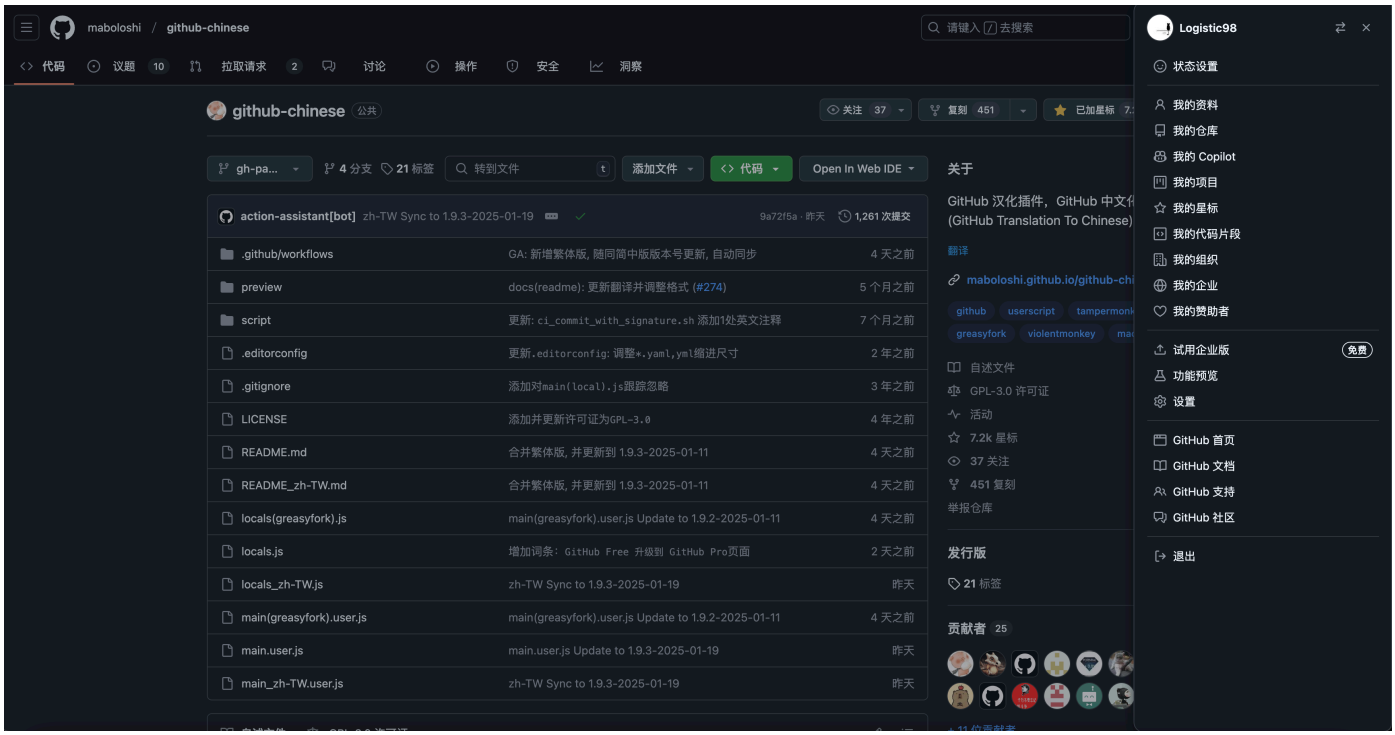


4、code2prompt（开源）

Code2prompt是一个终端工具，能将代码库转化为单一的LLM提示，结合源码树结构，模板定制，以及令牌计数。

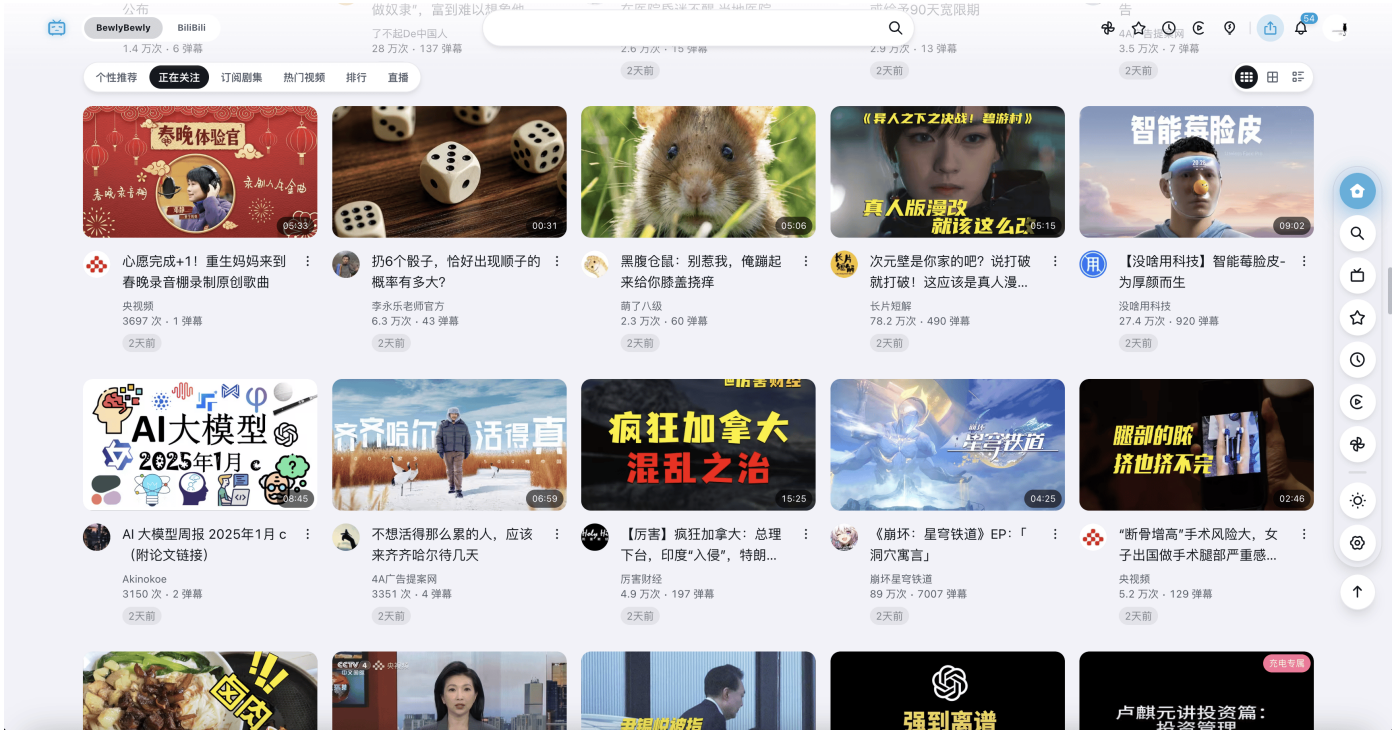
5、github-chinese（开源）

GitHub汉化的浏览器插件脚本，需要借助Tampermonkey去使用。[GitHub 中文化插件 - GreasyFork 托管 - 发布版]



6、[BewlyBewly]（开源）

BewlyBewly 是一个用于 BiliBili 的浏览器扩展，旨在通过重新设计 BiliBili 用户界面来提升用户体验。设计灵感来自于 YouTube、Vision OS 和 iOS，从而实现了更具视觉吸引力和用户友好性的界面。[Chrome插件安装地址]



7、[Audiblez]（开源）

这个工具可以将 Epub 电子书转成有声书，支持中文。

8、[Music Tag Web]（开源）

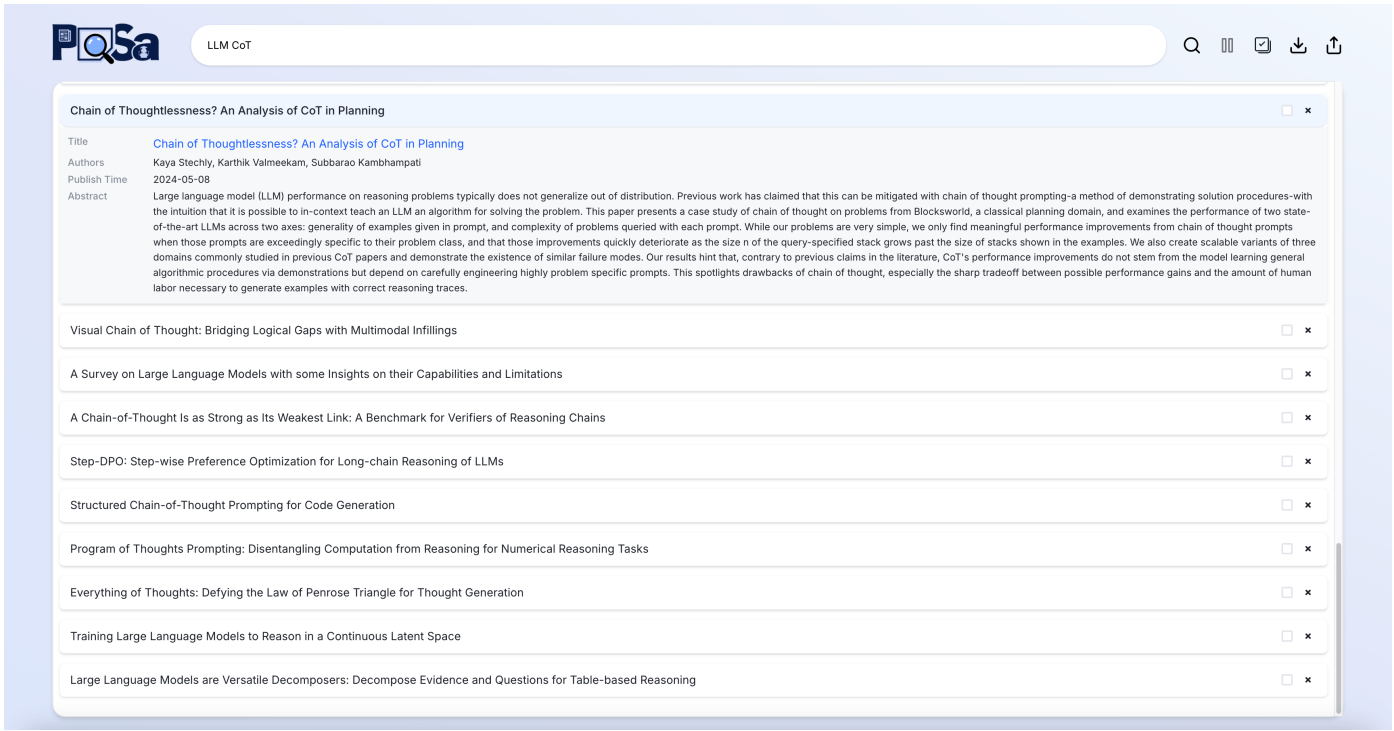
一款可以编辑歌曲的标题，专辑，艺术家，歌词，封面等元数据信息的Web工具。

9、DrissionPage（开源）

DrissionPage是一个基于Python的网页自动化工具，结合了浏览器自动化的便利性和requests库的高效率。它提供了三种页面对象：ChromiumPage、WebPage和SessionPage，分别适用于不同的使用场景，帮助开发者高效完成网页自动化任务。官方文档

10、PaSa（开源）

PaSa是字节跳动开源的由大型语言模型支持的高级论文搜索Agent。它可以自主地做出一系列决策，包括调用搜索工具、阅读论文、选择相关参考文献等，最终为复杂的学术查询获得全面准确的结果。官方试用



11、sshfs（开源）

这是一个基于 SFTP 协议的文件系统工具，可通过 SSH 协议将远程文件系统挂载到本地。它操作简单，仅需一条命令，即可像访问本地文件系统一样管理远程文件和目录，兼容 Linux、BSD 和 macOS 系统。

12、FileCodeBox（开源）

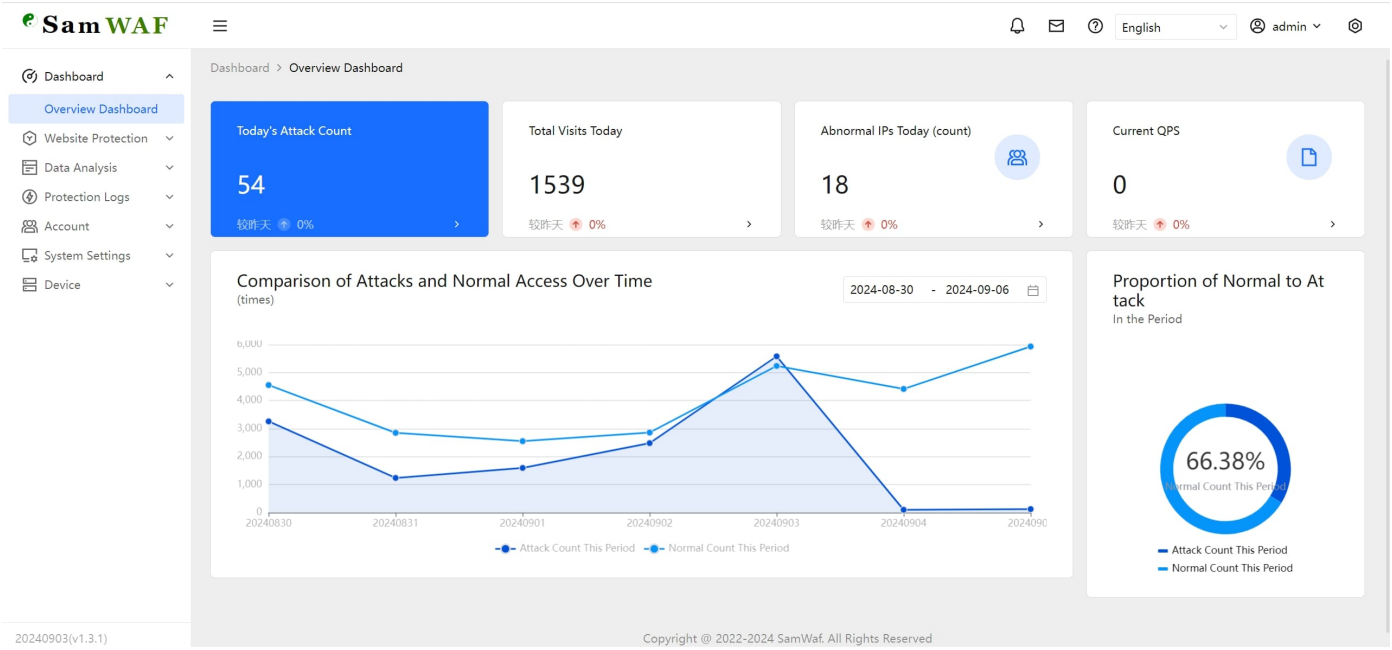
开源的文件快递柜工具，匿名口令分享文本，文件，像拿快递一样取文件。

13、upx（开源）

这是一款开源的可执行文件压缩工具，支持多种可执行文件格式（Windows、Linux、macOS）。它拥有出色的压缩比（50-70%），压缩后的文件可直接运行，适用于程序分发和大规模存储的场景。

14、SamWaf（开源）

一款完全开源的轻量级 Web 应用防火墙，支持私有化部署，提供 Bot 检测、URL 白名单、CC 防护、自定义防护规则等功能，适用于小型企业、工作室和个人网站。

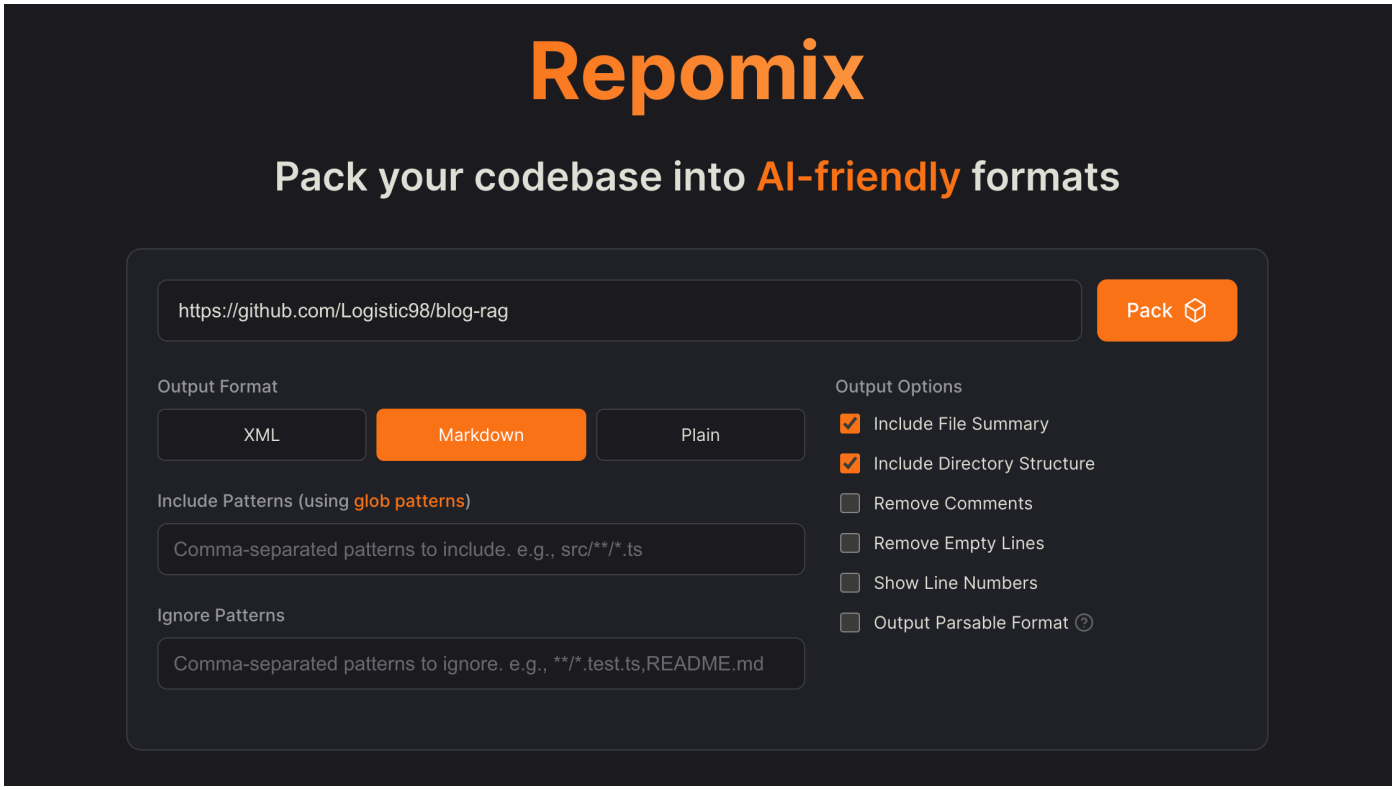


15、bunster（开源）

该项目是一个 Shell-to-Go 转译器（Transpiler），原理是先把 Shell 脚本转换为 Go 代码，然后利用 Go 工具链将其编译为二进制可执行文件，弥补了传统 Shell 脚本在性能、可移植性和安全性方面的不足。

16、Repomix（开源）

Repomix是一个专门用于将整个代码库打包成单一的、AI友好的文件。这个工具可以让开发者轻松地将他们的代码库提供给大语言模型进行分析和处理。官方试用

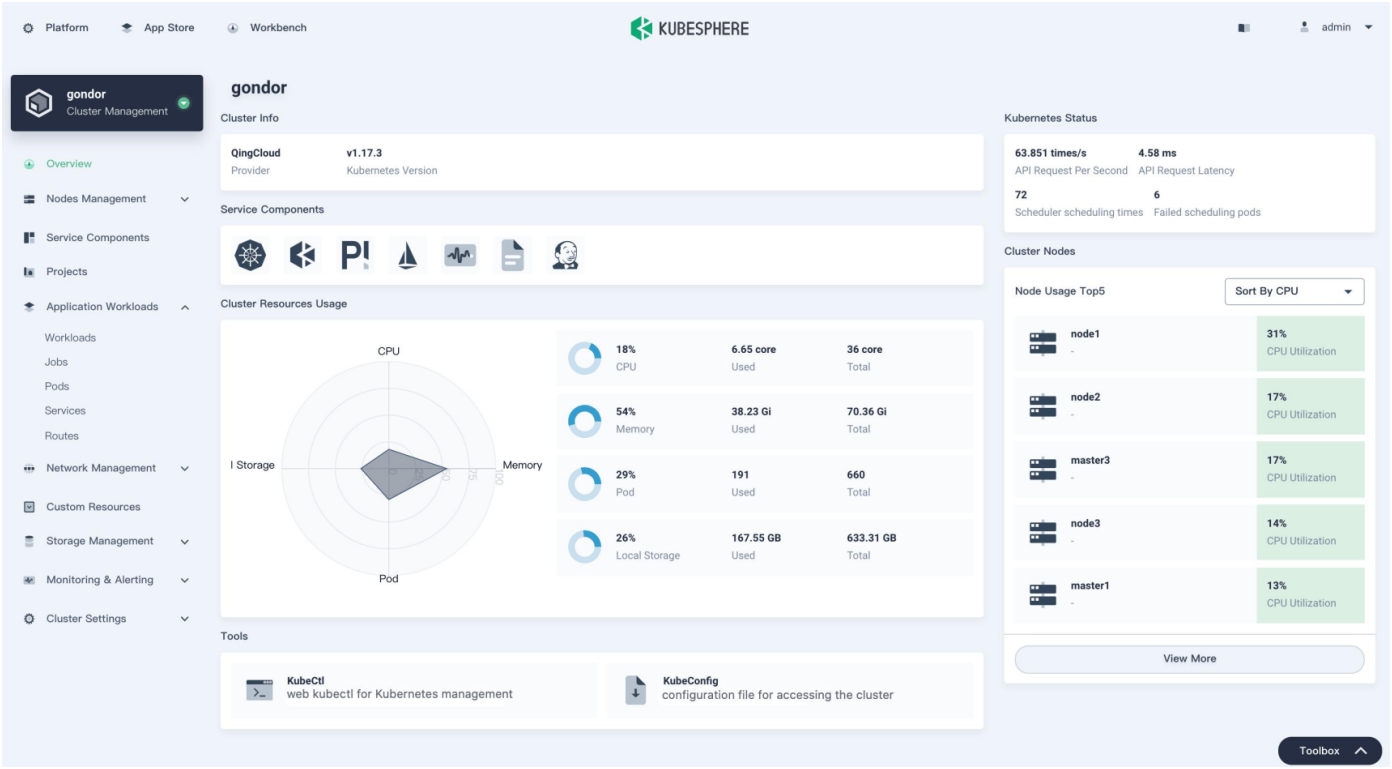


17、nginx-proxy（开源）

该项目可以自动为 Docker 容器提供 Nginx 反向代理服务。它能够实时监听 Docker 容器的启动和停止事件，自动为每个 Docker 容器配置 Nginx 反向代理，无需手动干预，极大简化了容器环境下的 Nginx 配置流程。

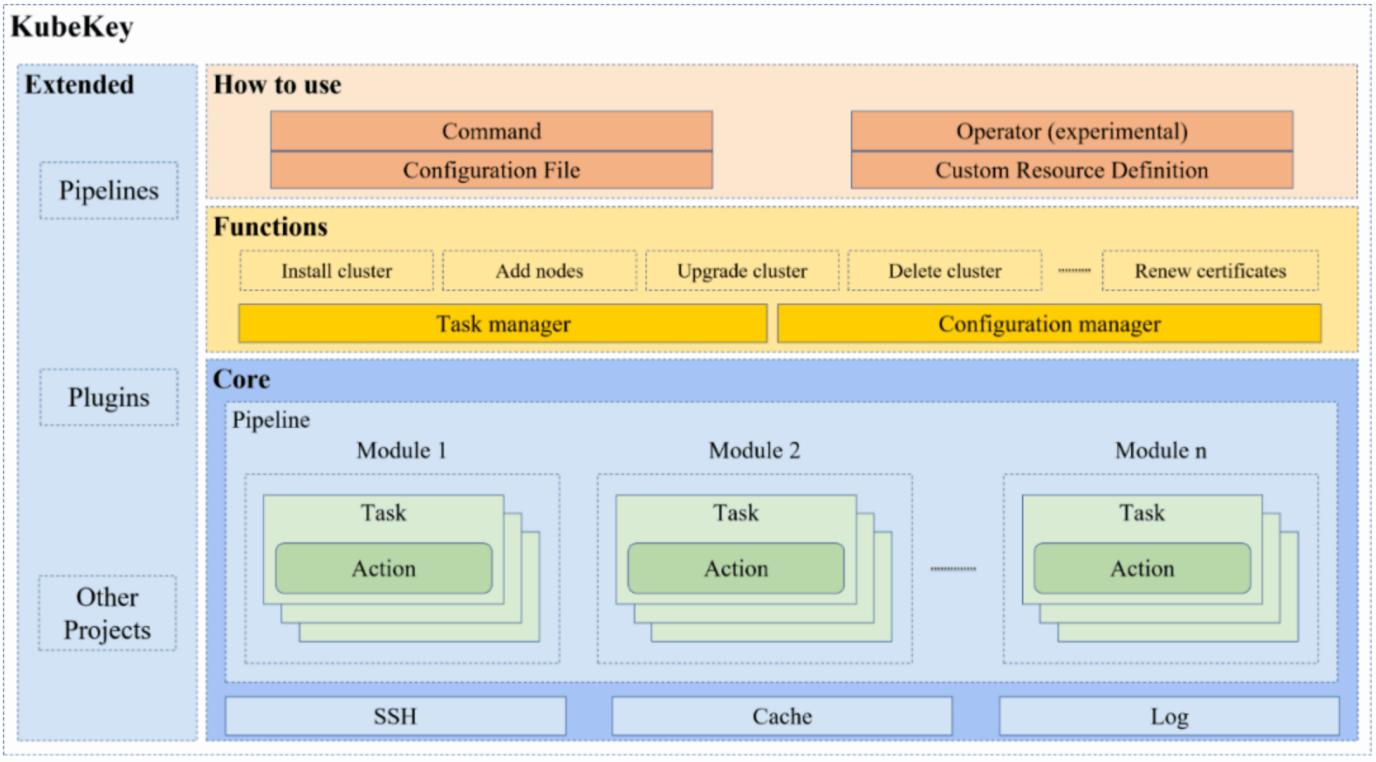
18、KubeSphere（开源）

KubeSphere是在Kubernetes之上构建的面向云原生应用的分布式操作系统，完全开源，支持多云与多集群管理，提供全栈的IT自动化运维能力，简化企业的DevOps工作流。它的架构可以非常方便地使第三方应用与云原生生态组件进行即插即用的集成。



19、KubeKey（开源）

KubeKey是基于Go语言开发的轻量级安装工具，它提供了一种灵活、快速、方便的方式来安装Kubernetes、Kubernetes和KubeSphere，以及相关的云原生附加组件，它也是扩展和升级集群的有效工具。此外，它还支持定制离线包，方便用户在离线环境下快速部署集群。

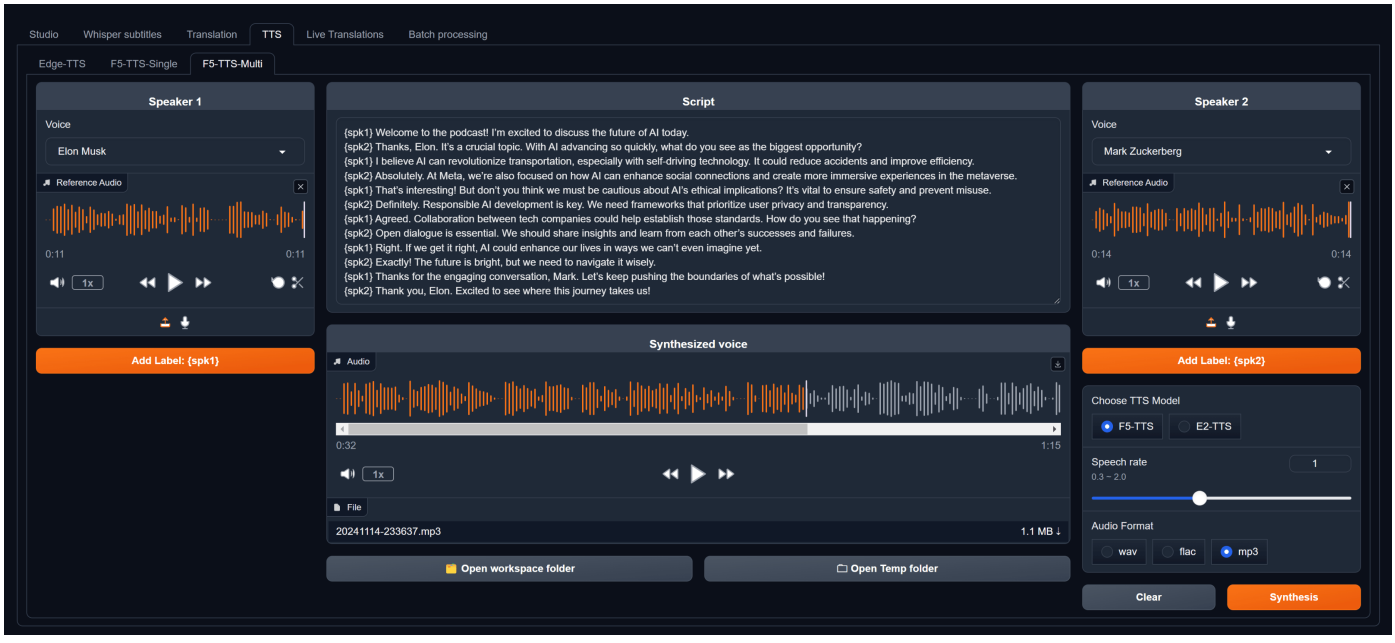


20、AigcPanel（开源）

AigcPanel是一个简单易用的一站式AI数字人系统，小白也可使用。 支持视频合成、声音合成、声音克隆，简化本地模型管理、一键导入和使用AI模型。

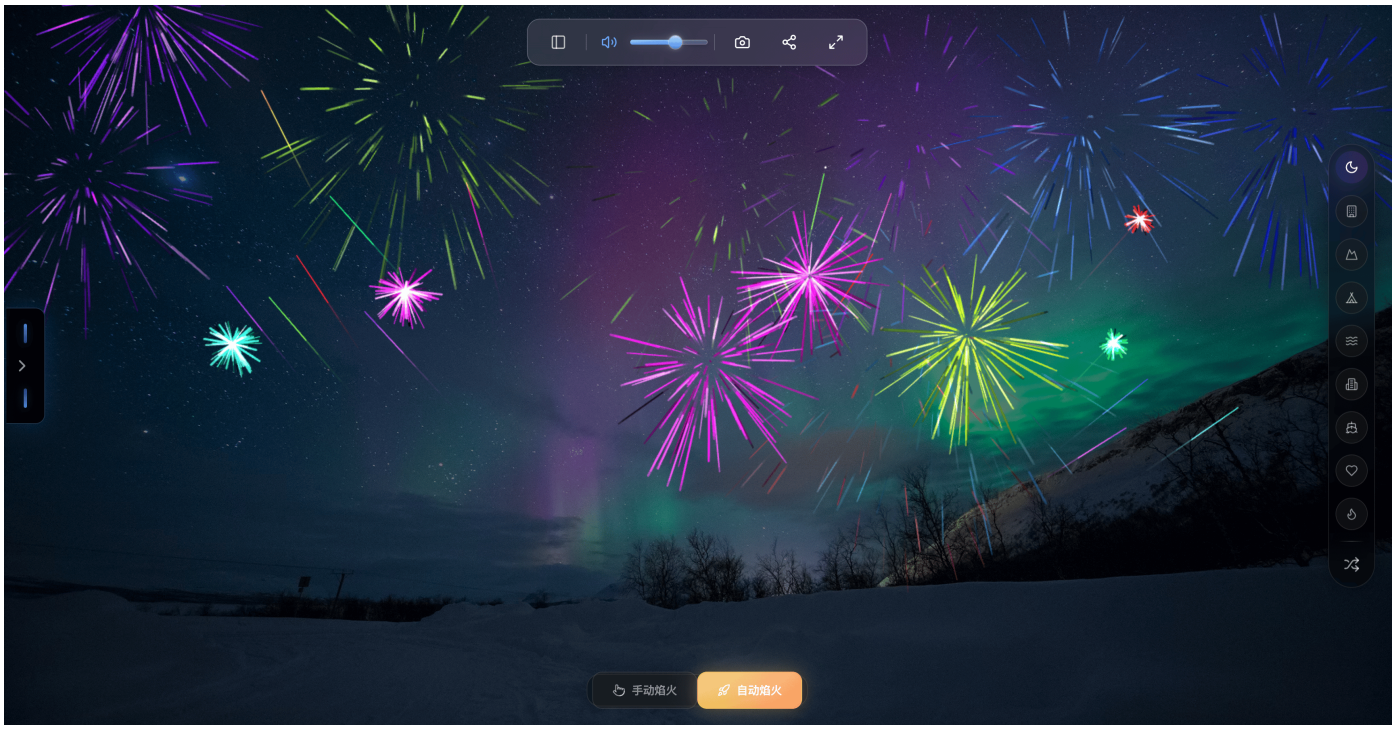
21、Voice-Pro（开源）

AI 语音工具，具有多种功能（语音识别、翻译、语音克隆、文本转语音），属于 Whisper 模型的 Web UI。



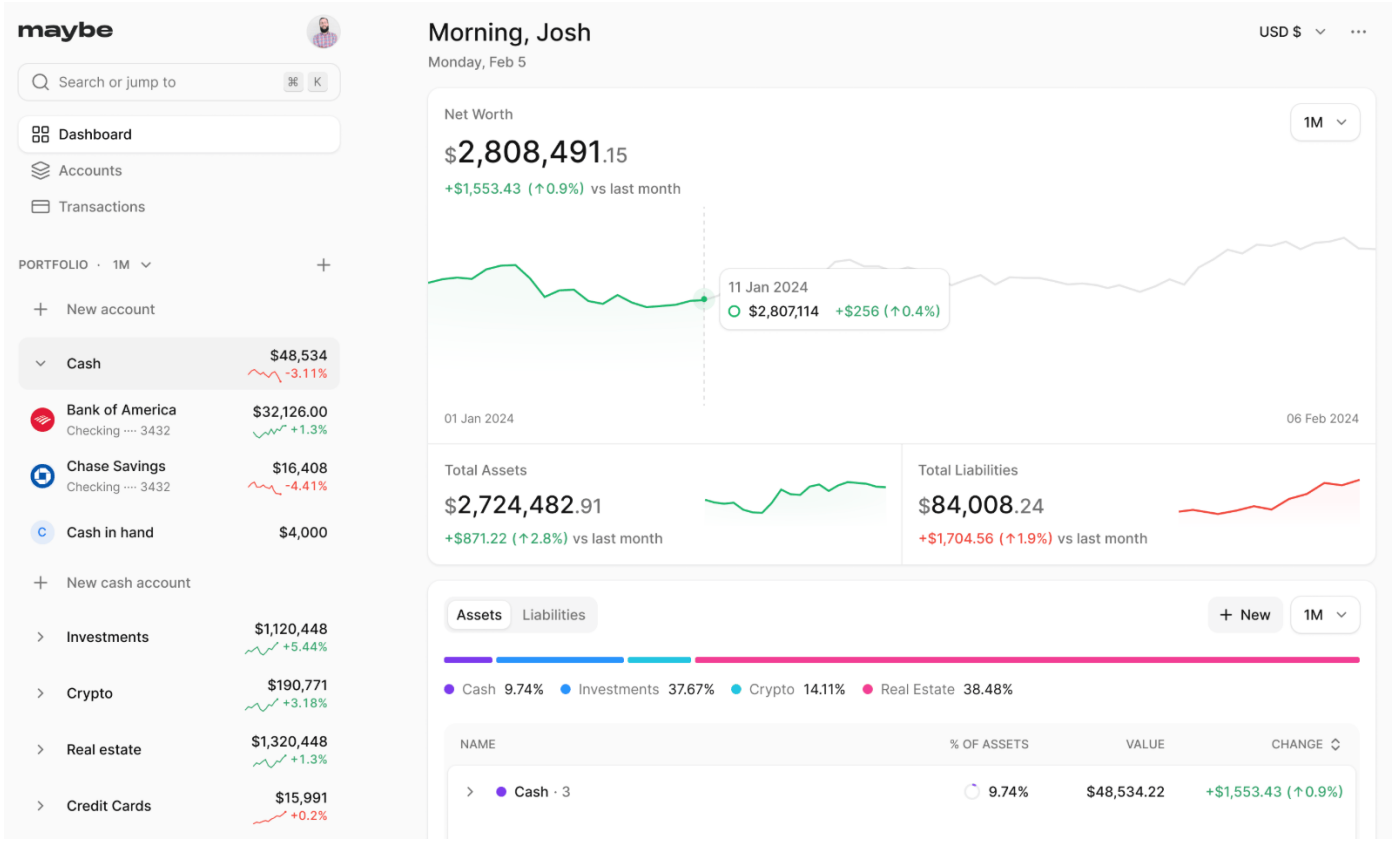
22、在线焰火模拟器（免费）

网页模拟焰火绽放效果的在线工具。



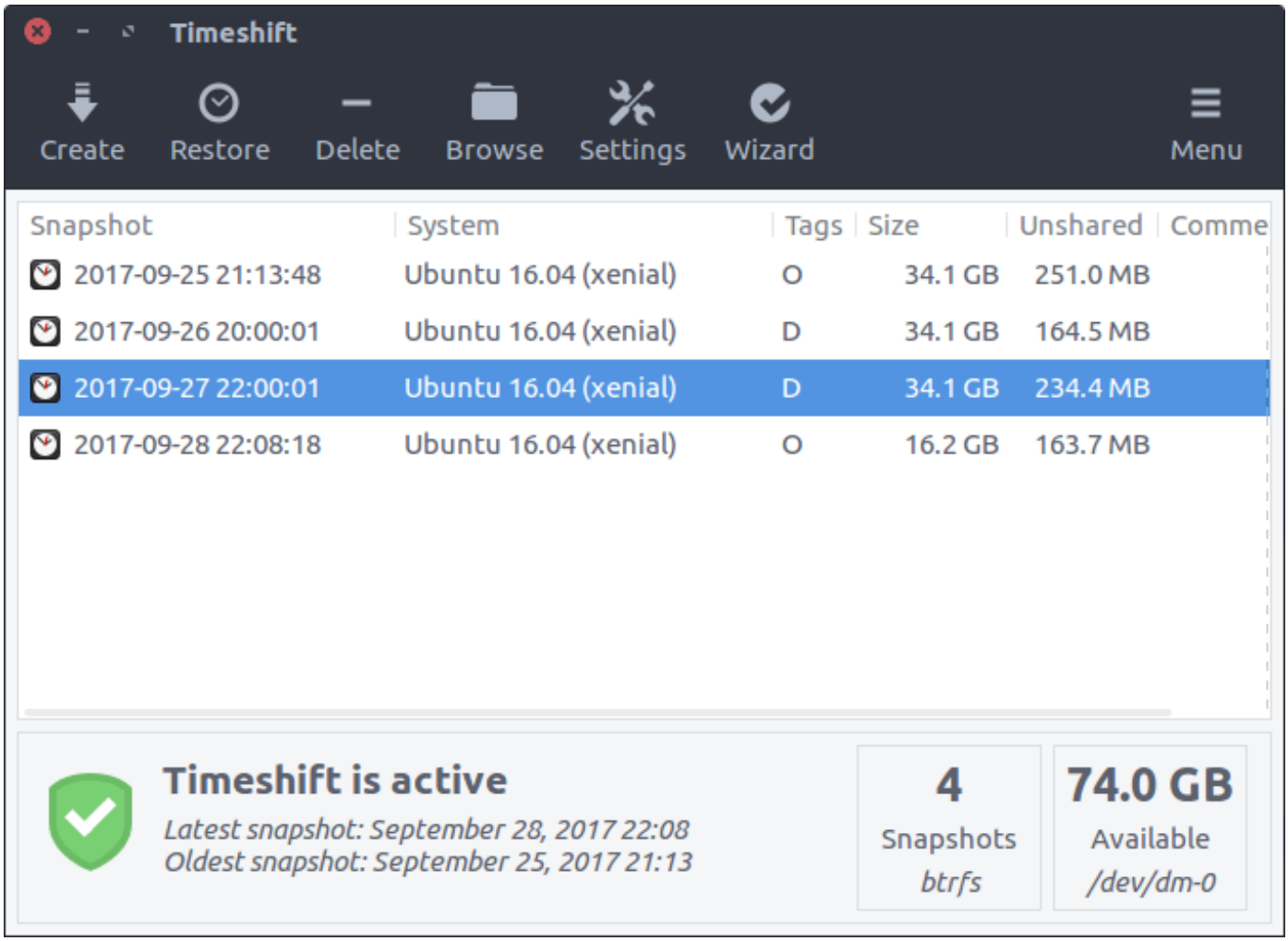
23、maybe（开源）

一个开源的金融应用项目，旨在提供一个灵活且易于扩展的金融服务平台。该项目支持多种金融操作，包括但不限于资产管理和交易处理。它的设计理念是模块化和可插拔，使得开发者可以轻松地集成和扩展功能。



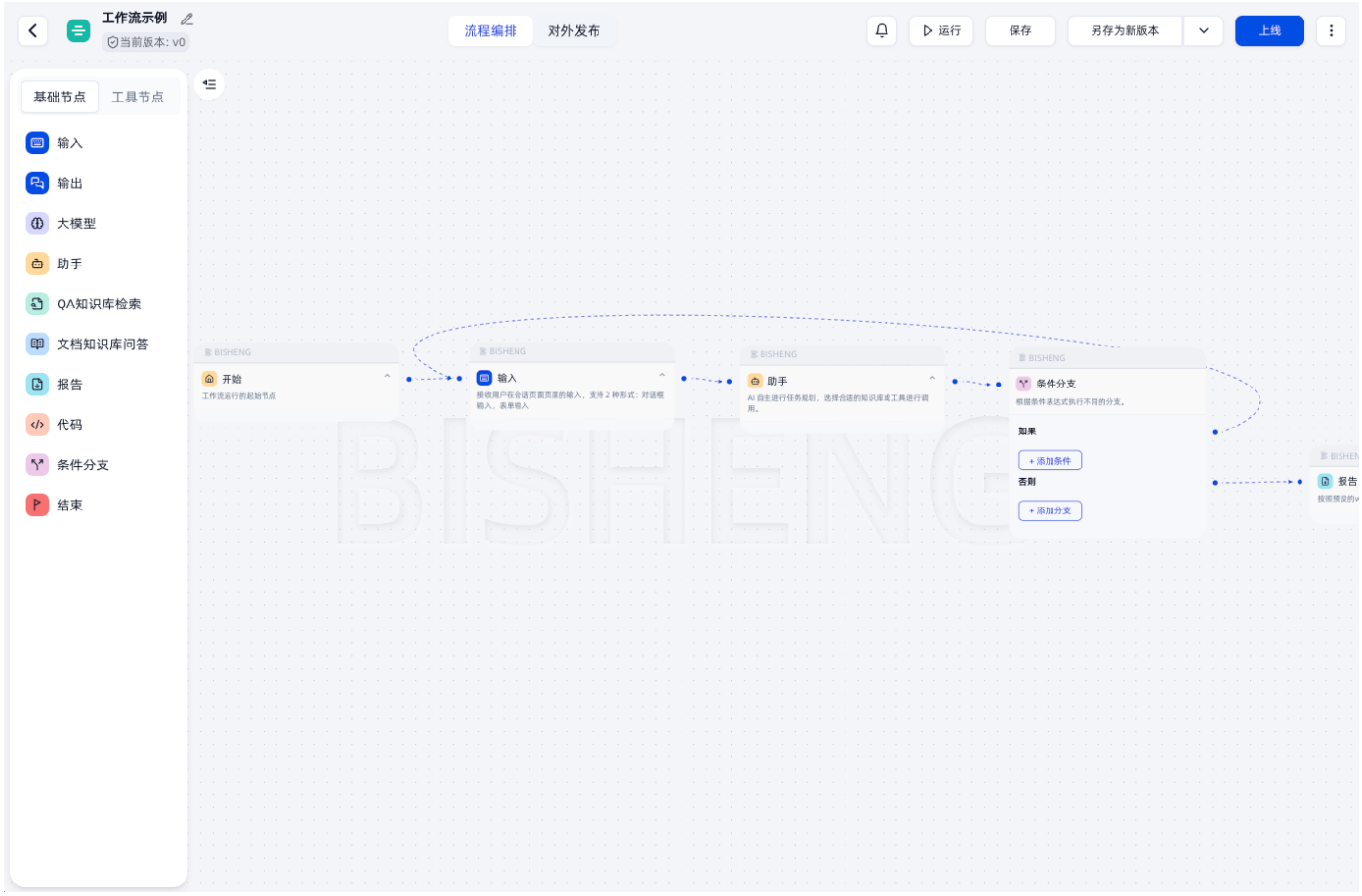
24、Timeshift（开源）

Linux 的时光机器，定期对文件系统生成增量快照，可以返回到指定时点。



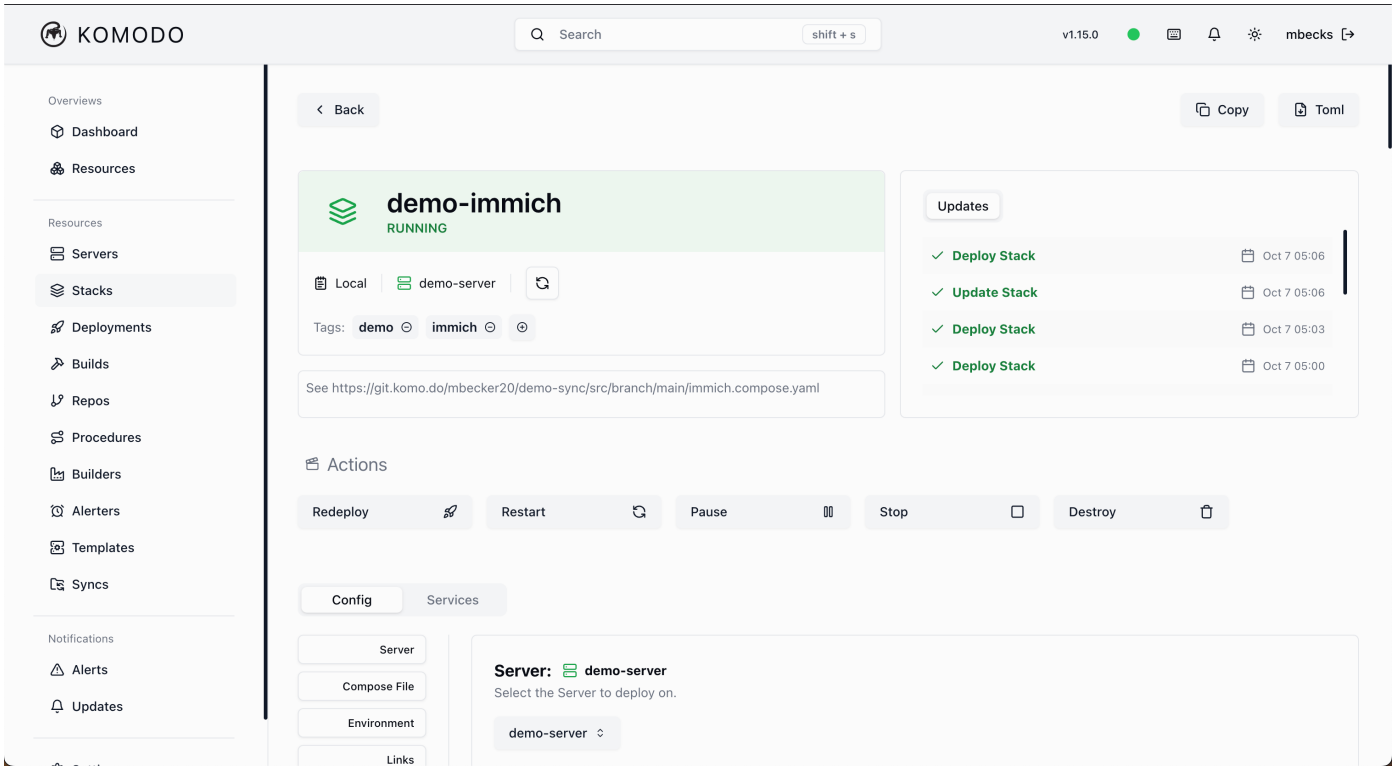
25、bisheng（开源）

bisheng 是一个开放的 LLM 应用 DevOps 平台，专注于企业场景，已被大量行业领先组织和财富 500 强企业使用。



26、komodo（开源）

一款开源、免费的多服务器部署平台，旨在帮助开发者在多个服务器上部署应用。它基于 Rust 和 TypeScript 构建，提供了一个界面简洁、灵活、无限制的自动化部署平台，支持无限扩展的服务器连接、管理 Docker 容器和环境变量等功能。



27、Deep Research Web UI

AI 深度研究的前端 UI，可以接入各种大模型，结合搜索引擎和网页抓取，对各种主题进行迭代式深入研究。

28、唐韵（免费）

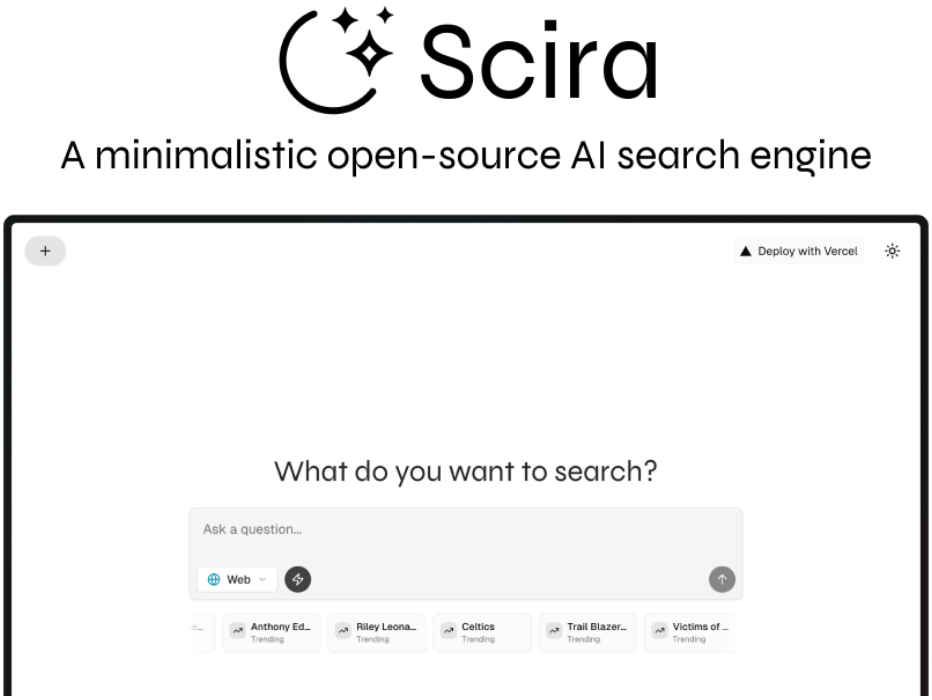
一款界面简洁的古诗词网站。

29、Amazon Kindle eBook Bulk Downloader（开源）

以更自动化的方式批量下载所有 Kindle 电子书，用于创建您已购买书籍的备份副本。

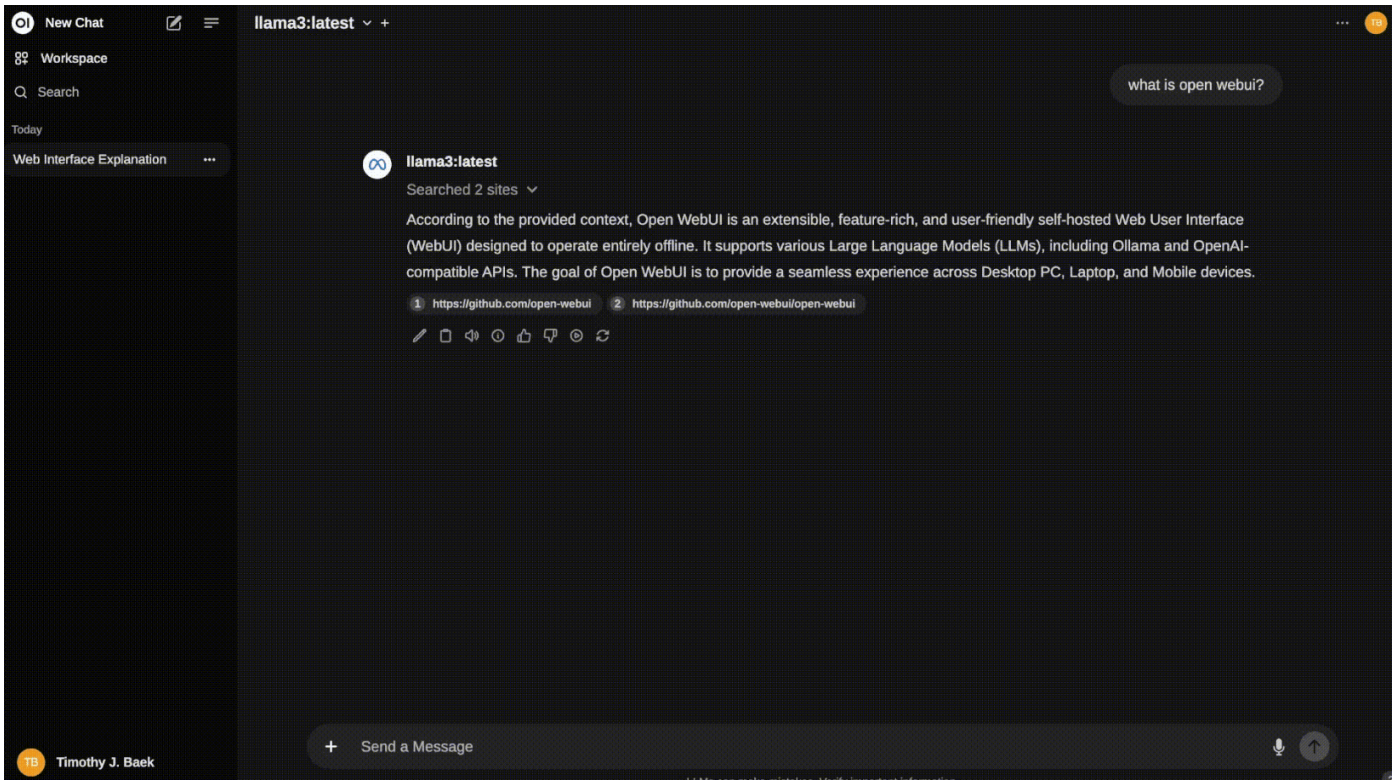
30、Scira（开源）

Scira是一款极简的 AI 搜索引擎，可帮助您在互联网上查找信息。



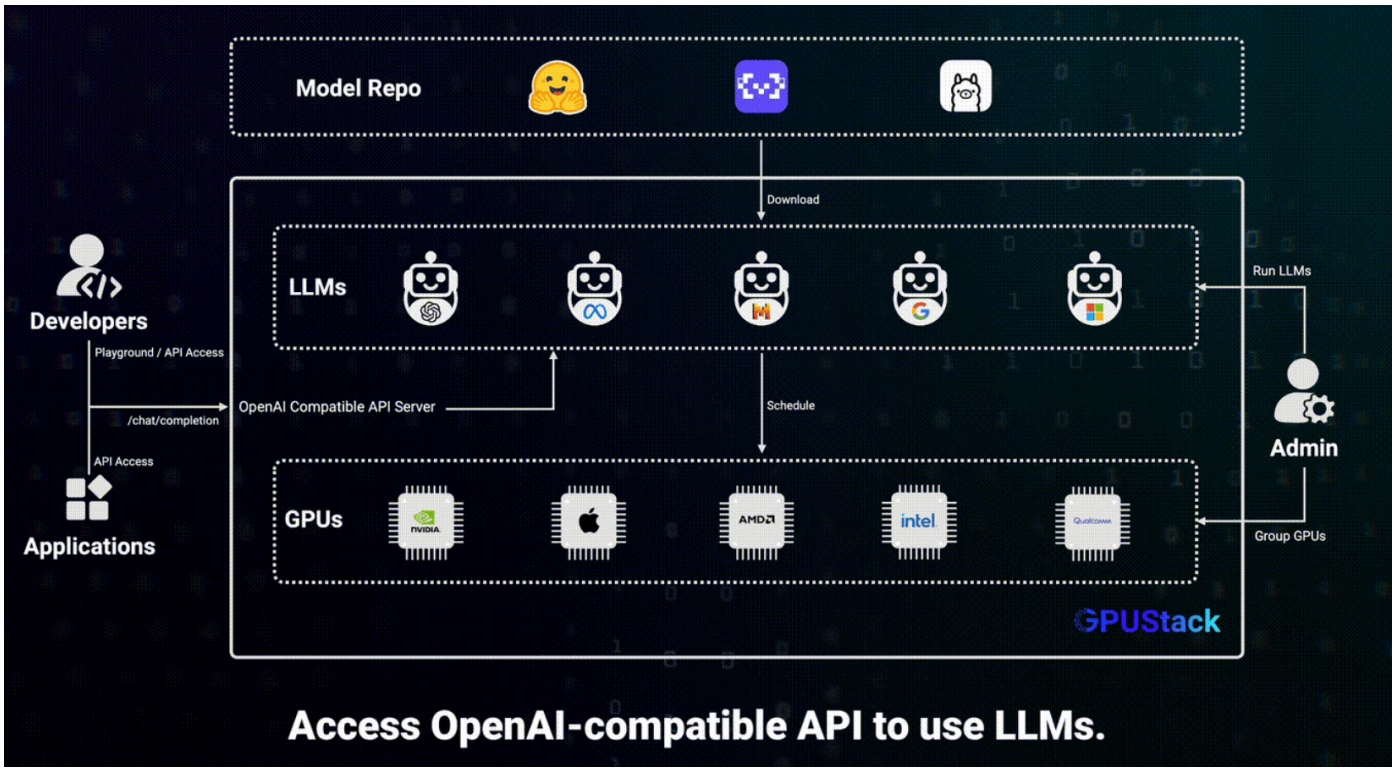
31、Open WebUI（开源）

Open WebUI 是一个可扩展、功能丰富、用户友好的自托管 AI 平台，旨在完全离线运行。它支持各种 LLM 运行器（如Ollama）和与OpenAI 兼容的 API，并内置RAG推理引擎，使其成为强大的 AI 部署解决方案。



32、GPUStack（开源）

GPUStack 是一个用于运行 AI 模型的开源 GPU 集群管理器，支持管理 Apple Mac、Windows PC 和 Linux 服务器上不同品牌的GPU。



33、earlyoom（开源）

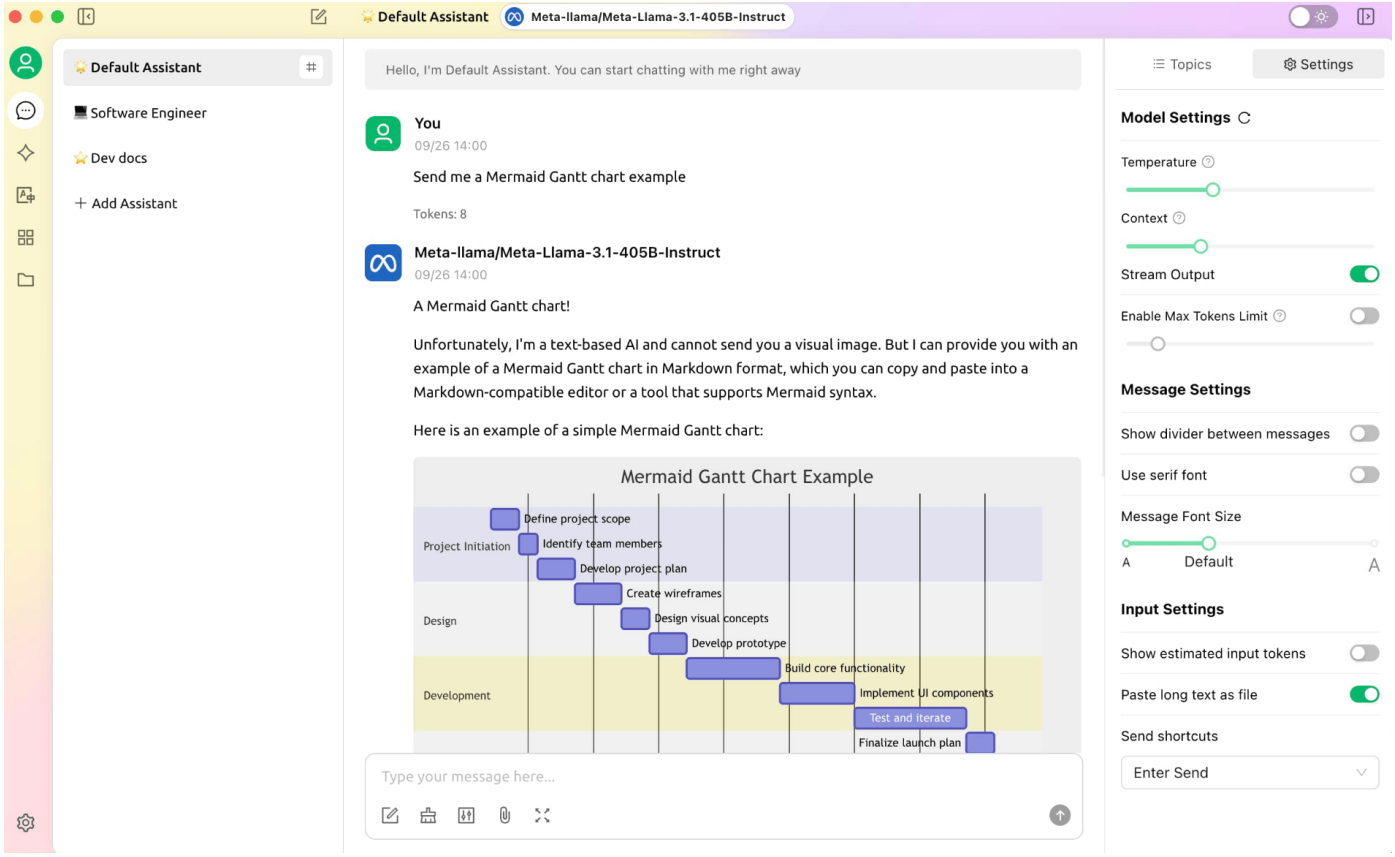
这是一款专为 Linux 设计的 OOM 守护进程，旨在弥补内核自带的 OOM Killer 仅在内存耗尽时才触发的不足。它能够提早干预（默认 10%），自动终止占用内存最多的进程，从而防止系统因内存耗尽而陷入卡死的状态。

34、ESP-Drone（开源）

这是一个基于 ESP32 的开源无人机项目，帮你制作一款低成本、可手机控制的小型无人机。

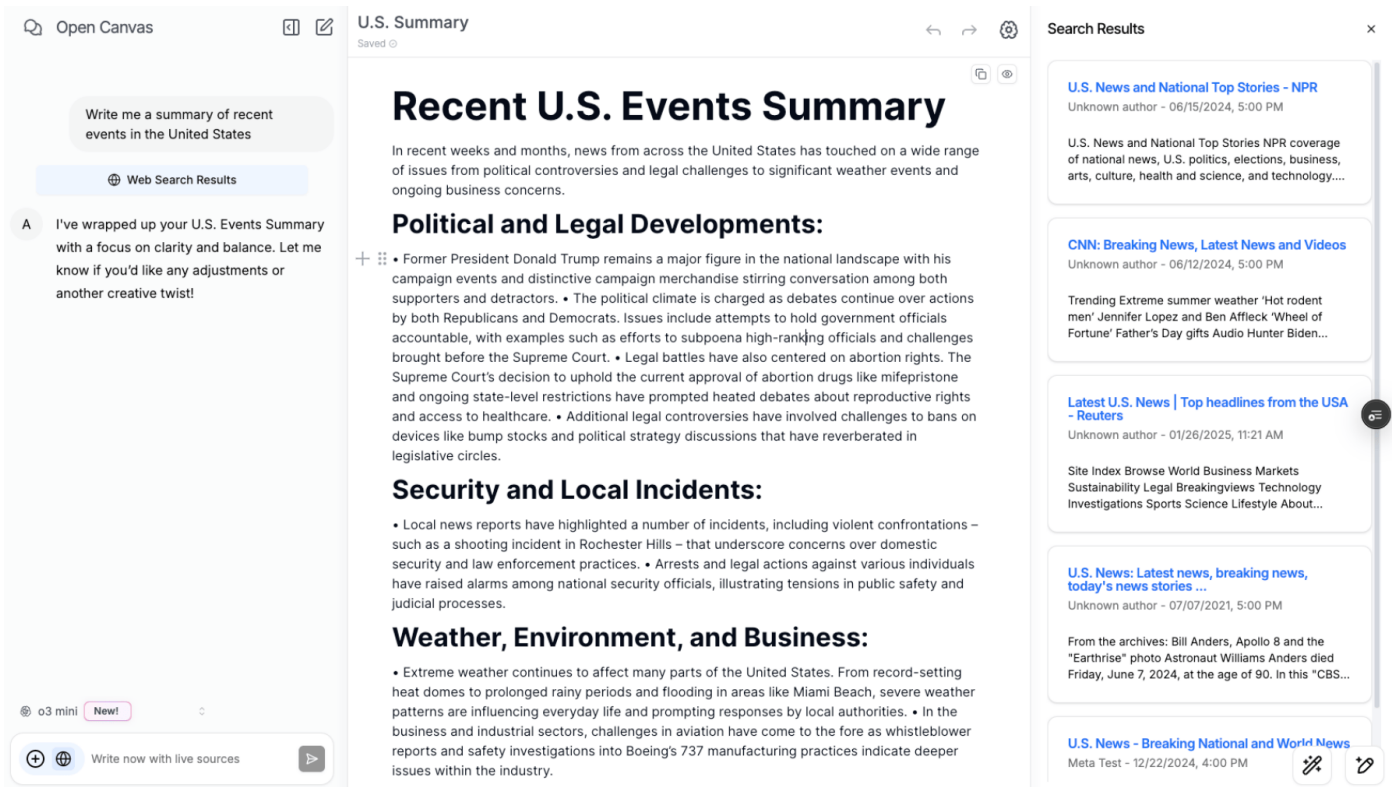
35、Cherry Studio（开源）

Cherry Studio 是一款支持多个大语言模型服务商的桌面客户端，兼容 Windows、Mac 和 Linux 系统。



36、Open Canvas（开源）

Open Canvas 是一个由 LangChain 团队开发的应用程序，该工具的定位是多功能的，它旨在服务于文档编辑、内容创作和编程，同时提供强大的协作能力。

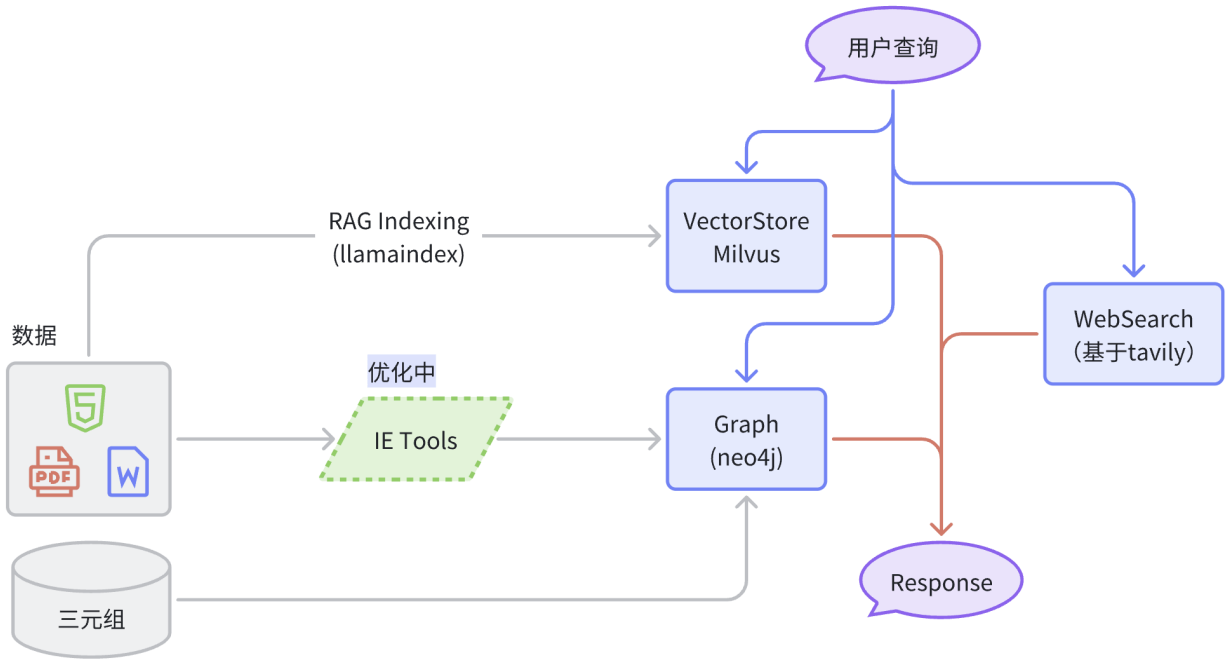


37、Clash Verge（开源）

Clash Verge是Clash内核的GUI图形客户端，支持Windows、Linux、macOS系统，分流规则功能强大且支持多种代理协议。Clash Verge 还在继续更新，而 Clash for Windows 已经停止更新了。

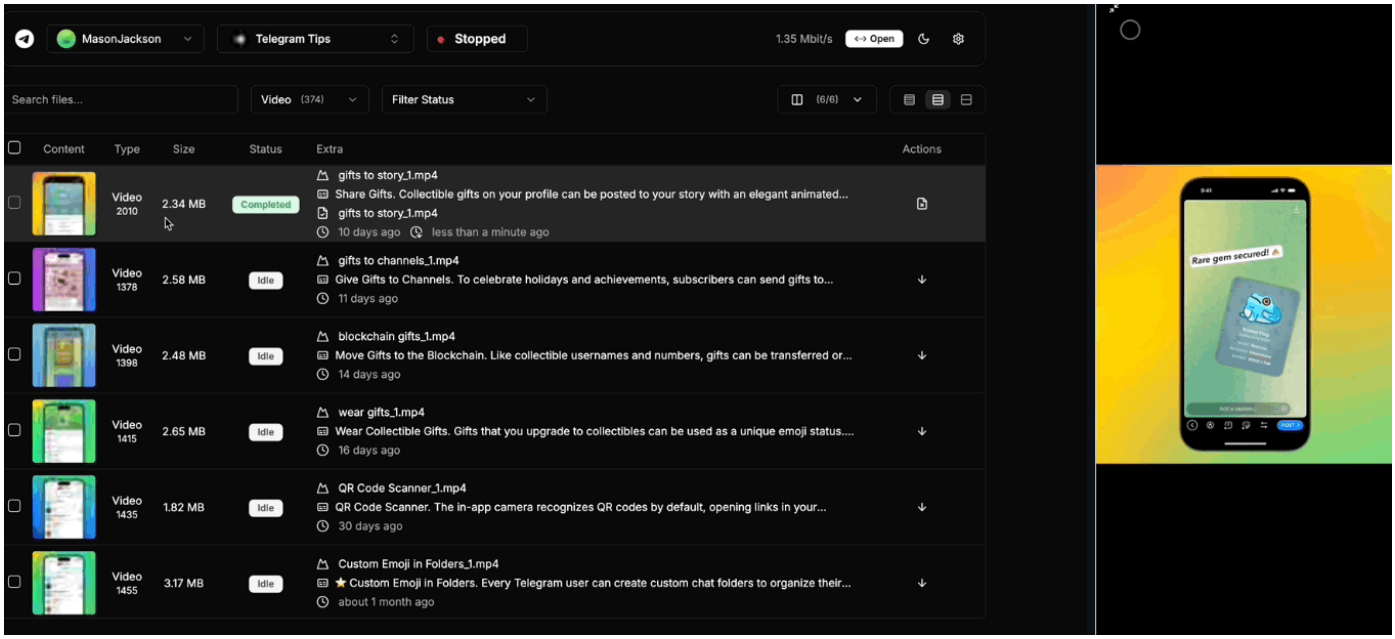
38、Yuxi-Know（开源）

语析是一个强大的问答平台，结合了大模型 RAG 知识库与知识图谱技术，基于 Llamaindex + VueJS + FastAPI + Neo4j 构建。



39、Telegram Files（开源）

开源的 Telegram 文件下载器，支持多频道、多账户同时下载。

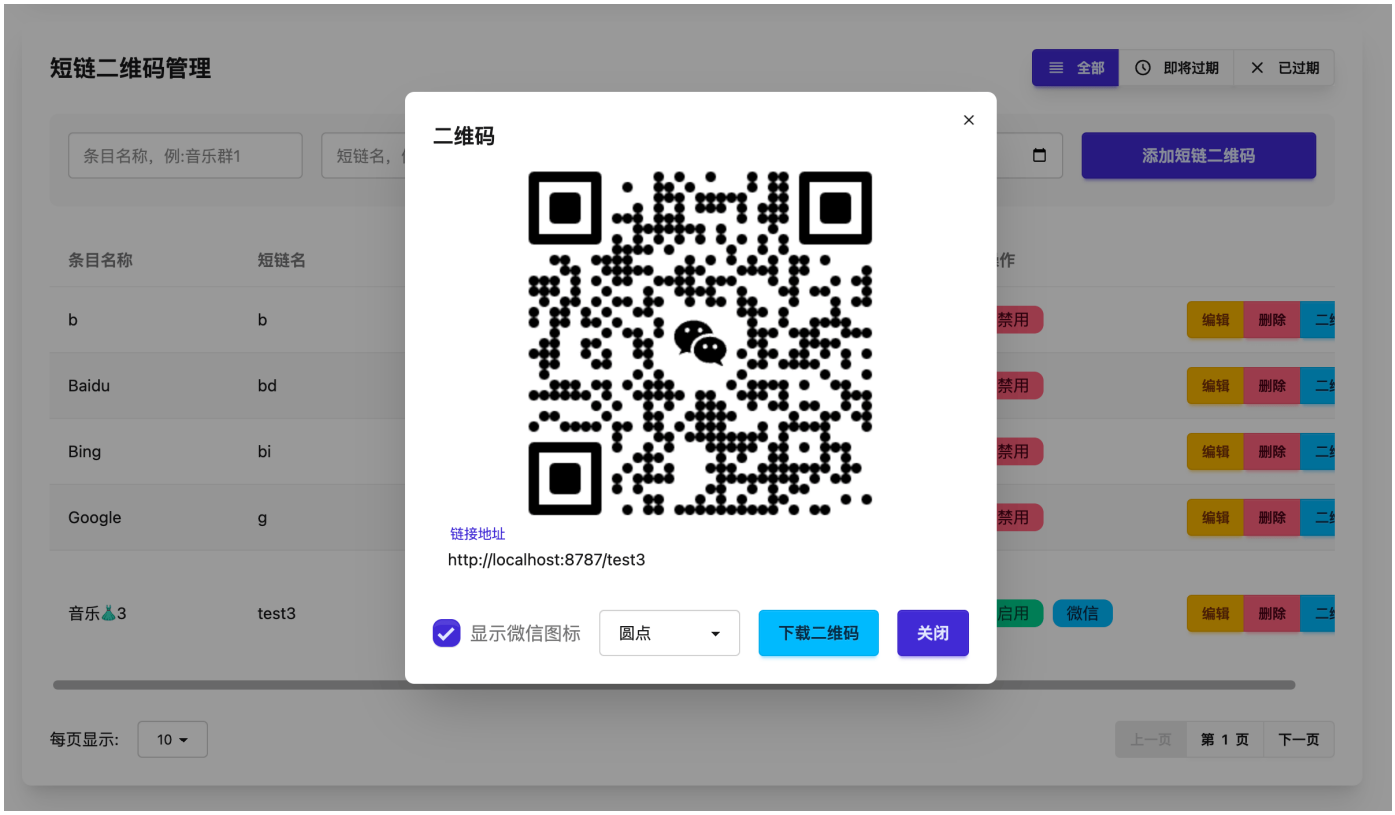


40、obsidian-cloud-sync（开源）

一个开源的 Obsidian 插件，将笔记自动同步到多种云盘服务。

41、serverless-qrcode-hub（开源）

微信群聊的二维码频繁变动，这个工具可以生成永久二维码，基于 Cloudflare Workers 和 KV 存储。

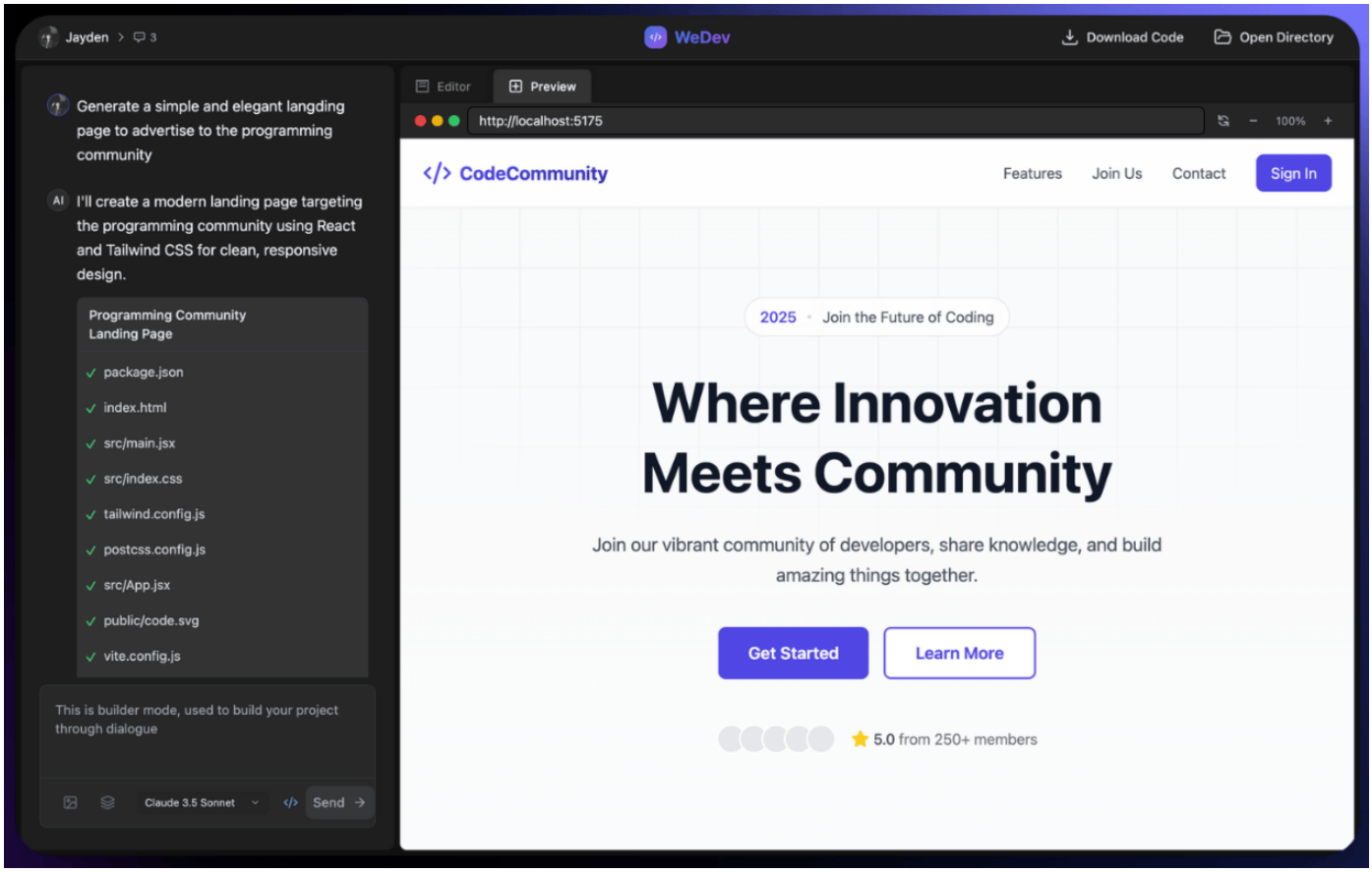


42、MarkPDFDown（开源）

基于大模型的 PDF 转 Markdown 工具，实现文档结构化转换。

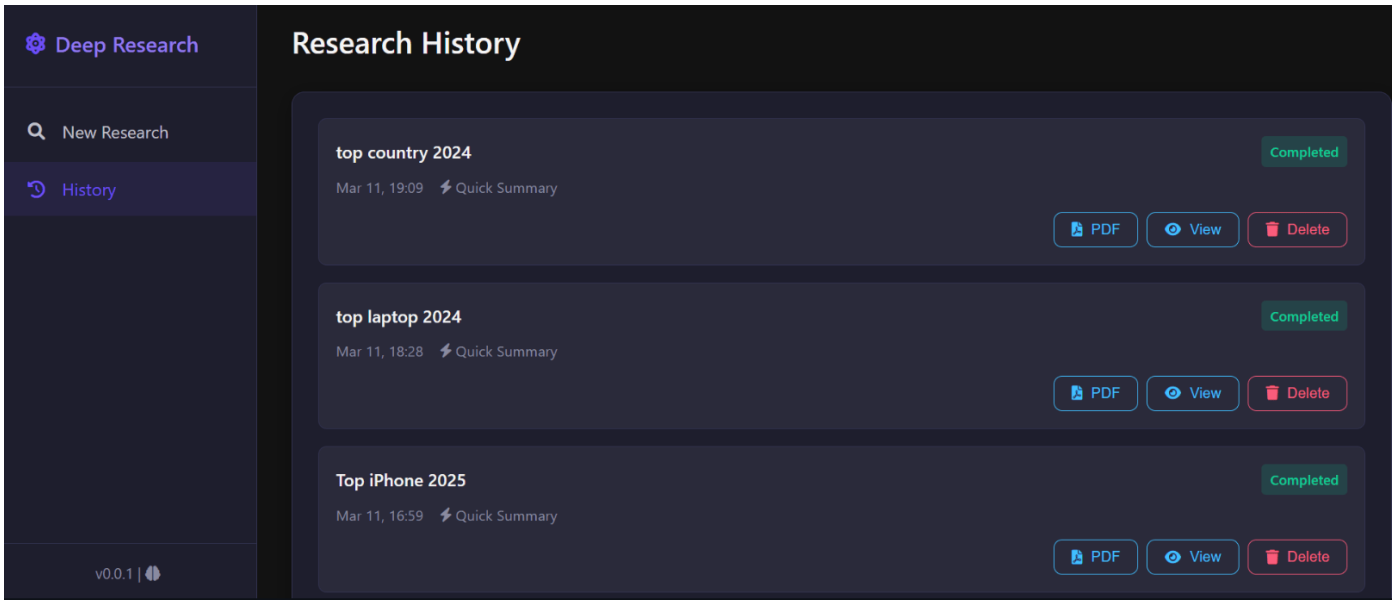
43、we0（开源）

通过 AI 生成应用程序，支持后端生成和前端生成，还可以 Sketch/Figma 设计稿1:1还原。官网试用



44、Local Deep Research（开源）

一款功能强大的人工智能深度研究助手，可使用多个LLM和网络搜索执行深度迭代分析。该系统可在本地运行以保护隐私，也可配置商业LLM API以增强功能。

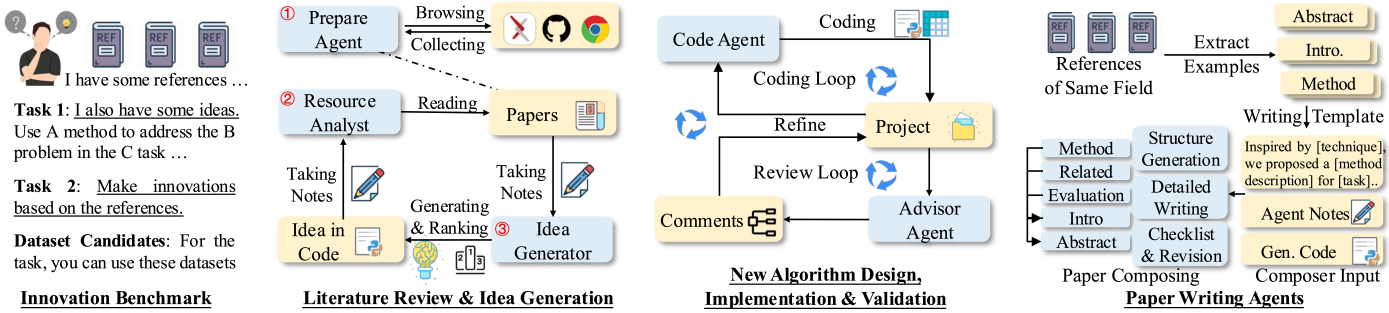


45、AI-Researcher（开源）

AI-Researcher 是香港大学数据科学实验室推出的开源自动化科学研究工具，基于LLM Agent 实现从研究想法到论文发表的全流程自动化，它支持用户在两种模式下操作：

- 一是提供详细的研究想法描述，系统据此生成实现策略；
- 二是提供参考文献，系统自主生成创新想法实施。

平台集成文献综述、想法生成、算法设计与验证、结果分析和论文撰写等核心功能，支持多领域研究，基于开源的基准测试套件评估研究质量。



46、RD-Agent（开源）

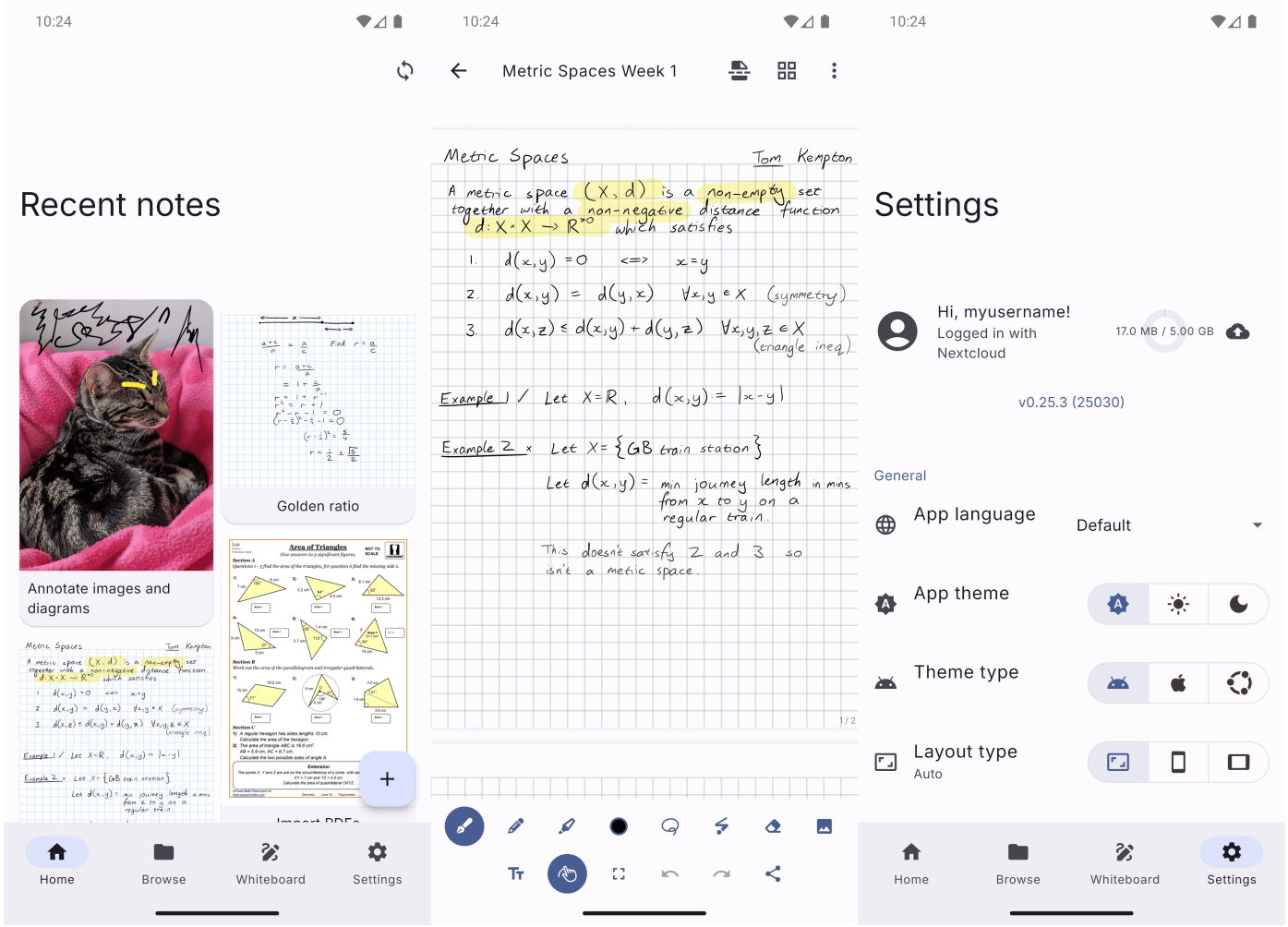
RD-Agent是由微软亚洲研究院推出的一款基于LLM的智能化工具。它将研究（Research）和开发（Development）两大模块无缝集成，形成一个持续反馈的自动化循环系统。RD-Agent通过自动化地生成假设、编写代码并回测结果，大幅提升研发效率和创新速度。

47、ivy（开源）

该项目可以将机器学习模型、工具和库从一个框架转换到另一个框架。开发者通过简单的函数即可完成代码的转换，支持 TensorFlow、PyTorch、JAX 等主流框架。

48、saber（开源）

这是一款开源的手写笔记应用，支持 Android、iOS、Windows、macOS、Linux 等平台。它提供夜间模式、多行公式高亮、密码保护等功能，适用于记录课堂笔记和整理工作思路等场景。



六、学习资源

1、[从零训练微型语言模型MiniMind](#)（中文）

从零开始训练小型语言模型，这不仅是一个微型语言模型的实现，更是一份入门 LLM 的教程，旨在降低学习和上手 LLM 的门槛。它提供了从数据预处理到模型训练、微调和推理的全流程代码和教程。最小模型仅 0.02B 参数，可在普通 GPU 上轻松运行。

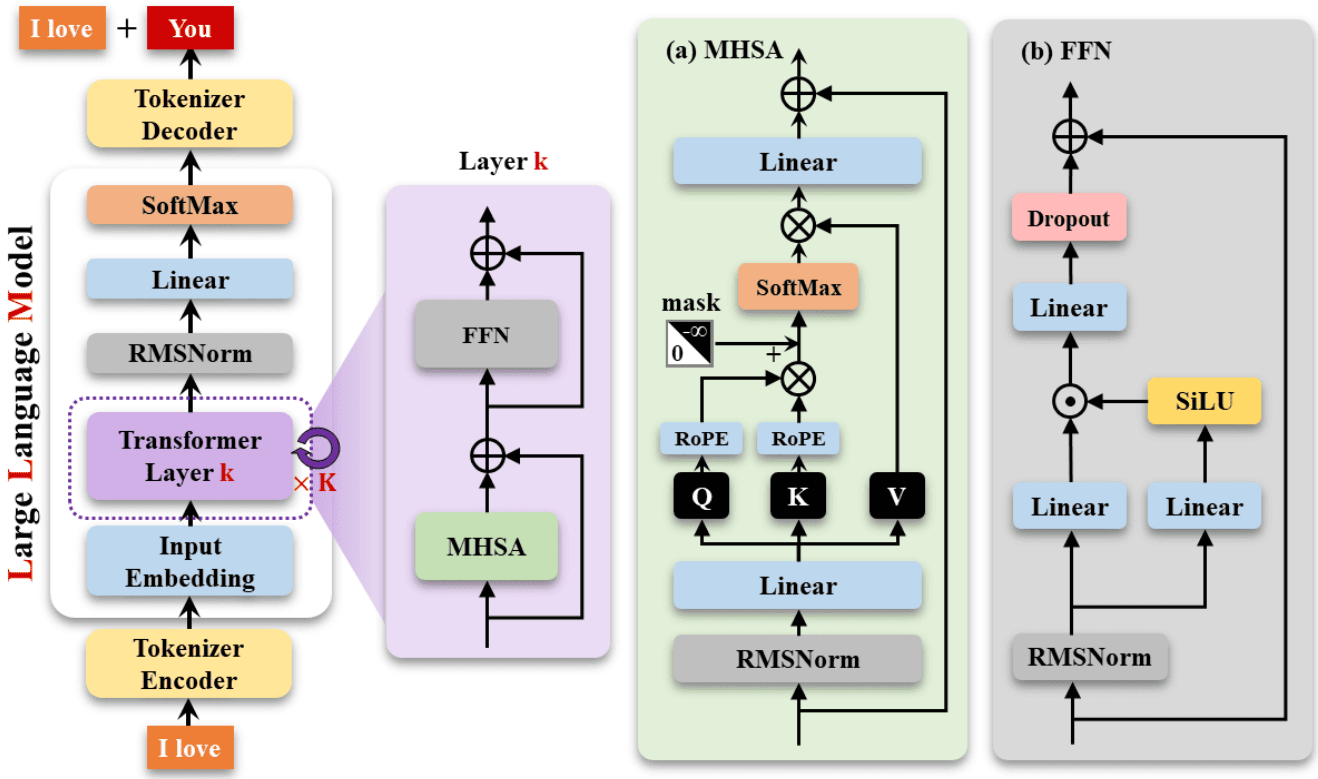


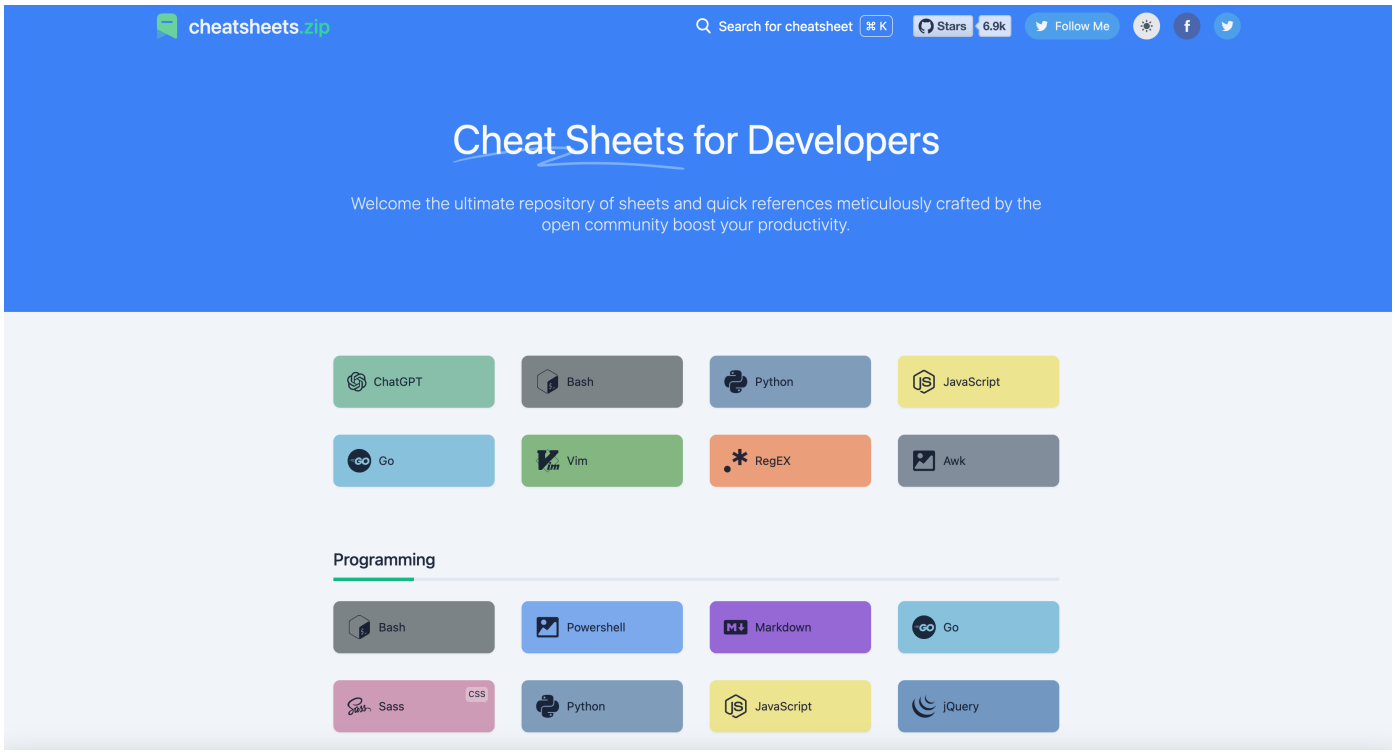
Fig. The Structure of MiniMind (Dense Model)

2、Foundations-of-LLMs（中文）

内含《大模型基础》开源书籍，该书是由浙江大学 DAILY 实验室开源的大语言模型教材，内容涵盖传统语言模型、大语言模型架构演化、Prompt 工程、参数高效微调、模型编辑、检索增强生成等方面。项目内还有LLM方向的经典论文、Arxiv前沿论文的收集。

3、Cheat Sheets for Developers（英文）

一份专为开发者准备的快速参考手册，旨在为开发者提供简洁、直观的速查表，内容涵盖多种编程语言、框架、Linux命令和数据库等。

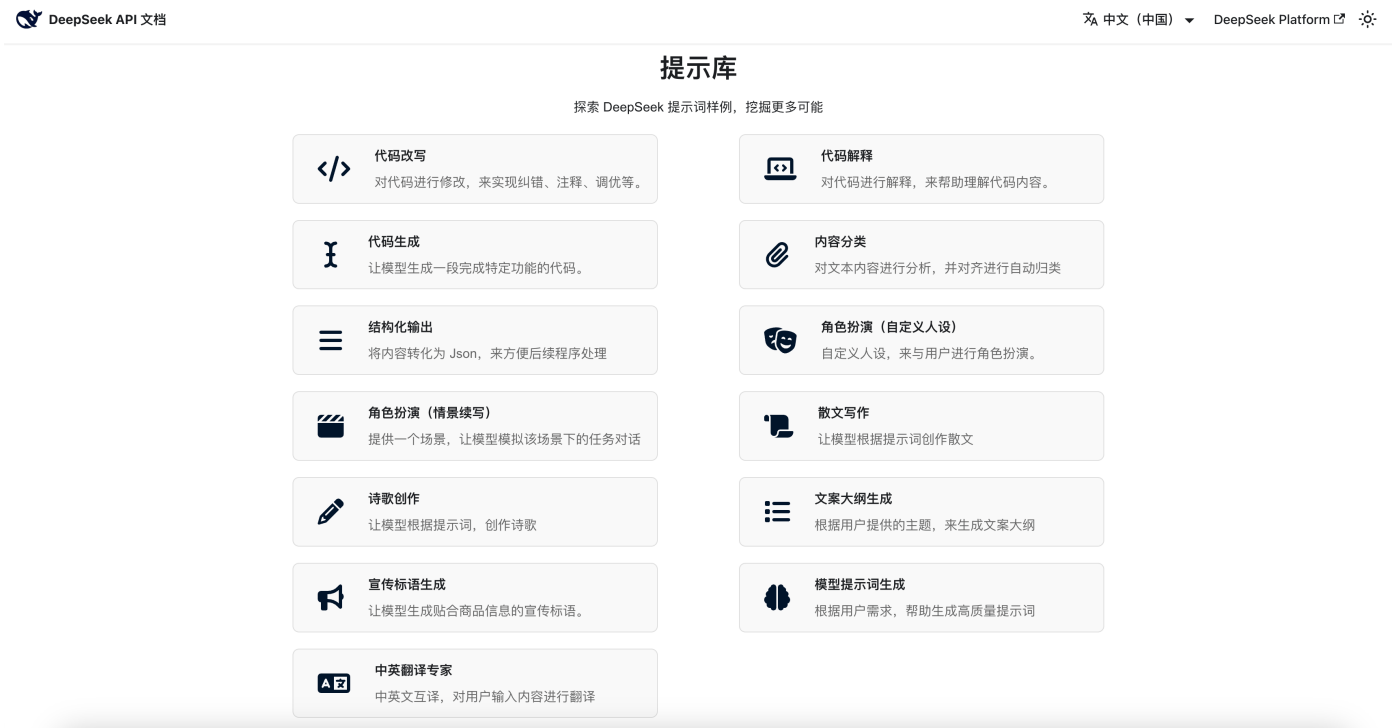


4、zh-style-guide（中文）

技术文档写作规范指南，旨在为中文技术文档的语言风格、结构样式、内容元素、标点符号、格式排版等方面提供参考规范。在线阅读

5、DeepSeek的提示库（中文）

DeepSeek官方文档里提供的一些特定场景需求的Prompt，可以参考它来写出高质量的提示词。

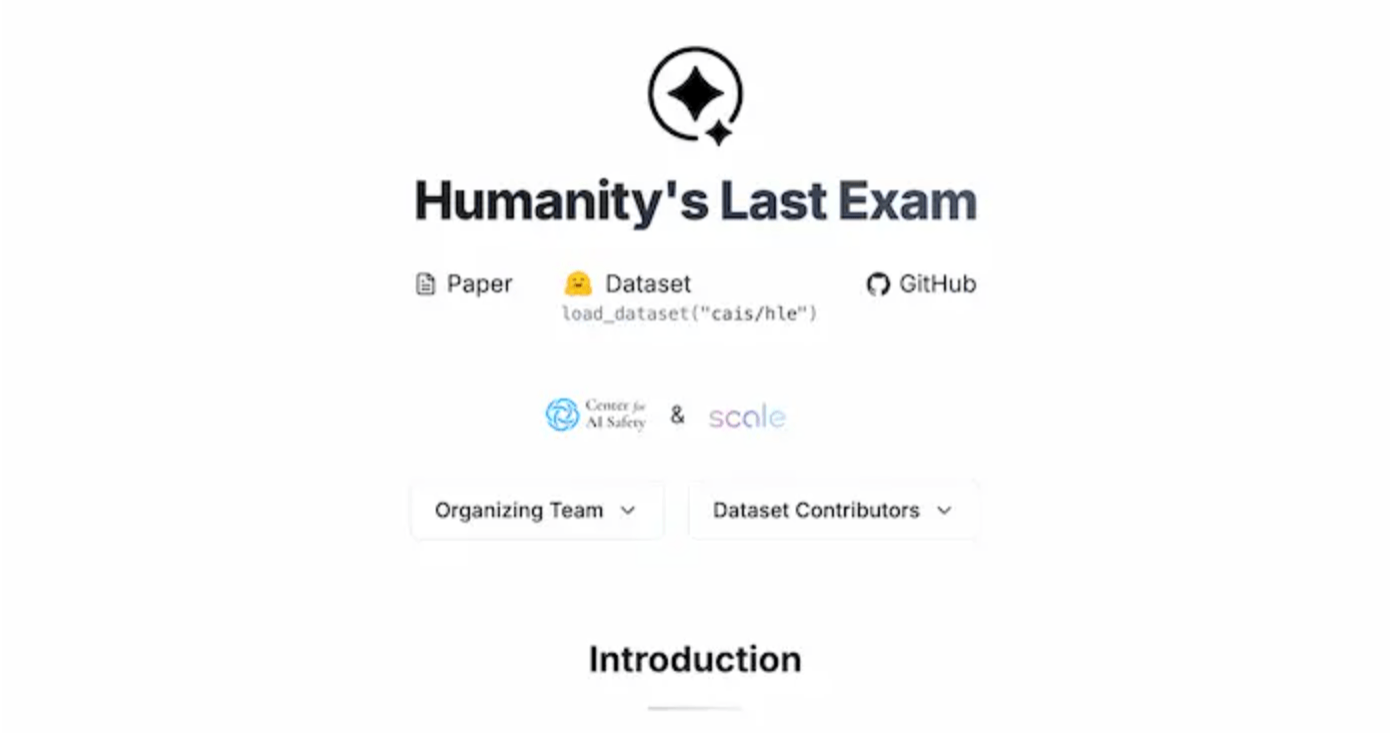


6、awesome-systematic-trading（中文）

一个精心整理的系统化交易资源列表，包括库、包、策略、书籍和教程。旨在帮助用户找到、开发和运行系统化交易（量化交易）策略所需的各种资源。

7、人类的最后考试（英文）

今年1月份，两家美国AI公司推出了一个测试集，包含3000道各种学科的题目。据他们说，只要AI模型通过了这个测试集，就表明AI智力已经超过了人类，也就是达到了AGI的水平，所以起名为“人类的最后考试”。截止2月3日，AI模型取得的最佳成绩是26.6%的正确率。

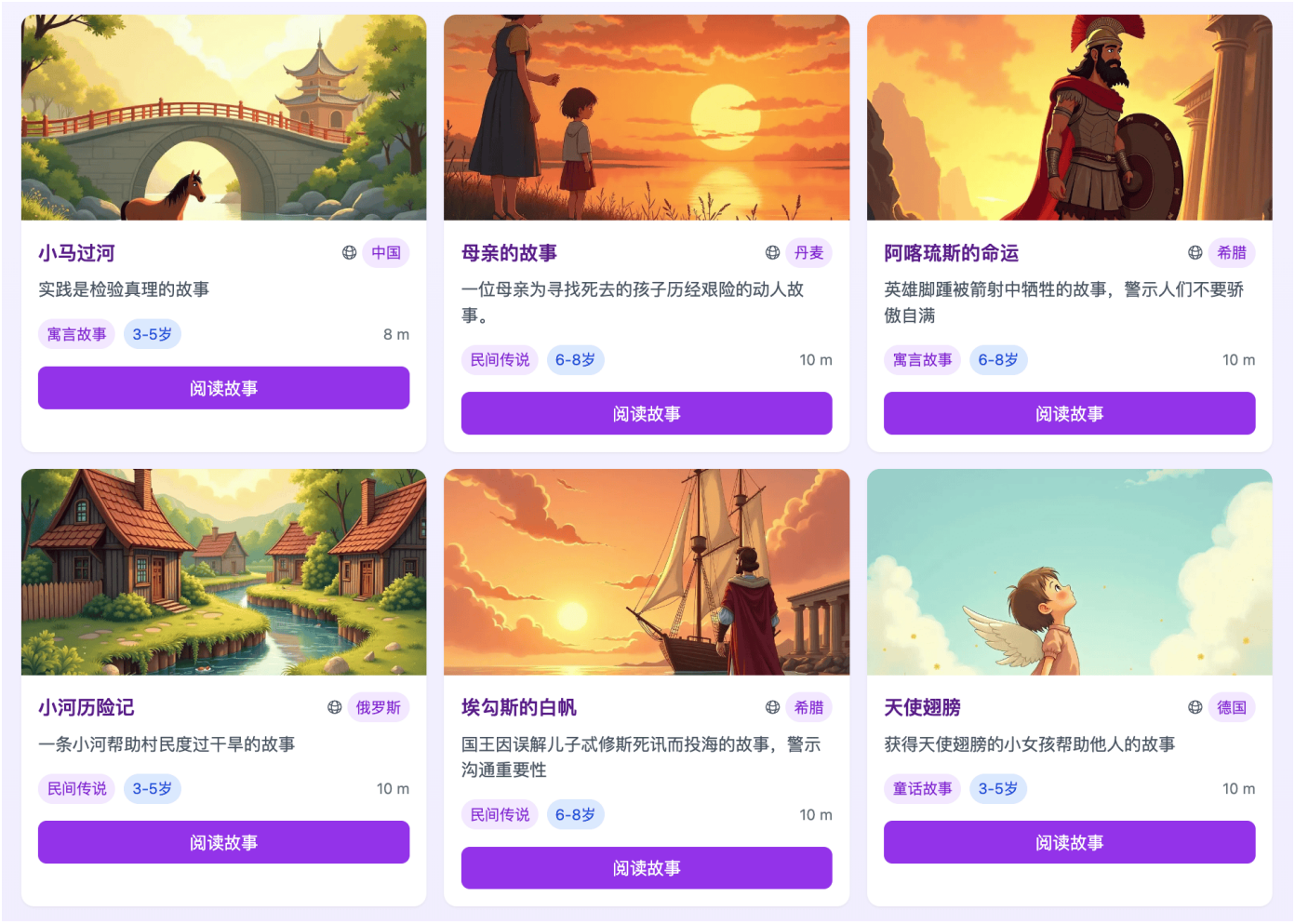


8、自洽的程序员（中文）

关于程序员如何管理情绪和职业心态的书籍，帮助程序员从负面情绪中解脱出来，更加坦然地面对自己的内心，从而实现“自洽”。

9、BeddyStories（中文）

一个儿童睡前故事网站，收集了全球经典的儿童睡前故事。



七、生活经验

1、DeepSeek实习经历及大模型择业思考

大公司的问题是，没有信仰，没有信心，人太杂。

- 首先是信仰，梁老板本人很有信仰，他相信AGI一定会达成，而且是有限的时间内可以达成，这个AGI一定是稀疏的，这份信仰传播给每一个下属。相比之下，字节领导对AGI的信仰不足，没有明确的判断，也不会把信仰传递到下属（人太多了，也传播不过来）。
- 第二个是信心，字节大部分情况下都在服务于OKR，业务最重要，如果对于业务没有明确好处的东西，大家会表现出没有信心，不愿意深挖。换言之，在字节很难抗住因为探索一件事情而没有产出的压力。我观察和我一起进字节的同学，做事非常仔细谨慎，一心一意服务业务，不敢越雷池半步。
- 第三个，人太杂，很难统一军心，很多时候别人做的东西自己也不关心，甚至会有人暗地使绊子。DeepSeek里则是大家有一种莫名的一致性，大家都是老板的粉丝，都信老板的话，平时谈话也都是自己哪里比友商强，哪里可以改进，整体氛围非常巩固军心，相互之间是促进关系。相比之下，我在字节看到了一些组的合并，合并的过程是先赛马再强行整合，大家相互敌对，浪费人力，浪费时间，浪费信仰，浪费资源，最后两败俱伤，把原本可以有领先机会的东西再次搞落后。

2、亚马逊河为什么没有桥?

南美洲的亚马逊河是世界第二长的河流，仅次于非洲的尼罗河。但是，这条河却是唯一一条没有任何桥梁的世界主要河流，原因如下：

- 首先，每年的雨季，亚马逊河都会泛滥，河流的宽度会从旱季的5公里变成50公里，很难造桥，变成下图所示的样子。

- 其次，亚马逊河沿岸人烟稀少，只有很少几个城镇。最大一个城市的人口只有50万，不存在前往河对岸的强烈需求。
- 最后，亚马逊河两岸都是原始森林，并没有现成道路。如果造桥就需要砍伐大量森林，修建引桥和公路，环境代价很大。

3、有利息的工作

银行存款有利息，存得越久，利息越多。工作也是一样，也有利息，如果今年的工作可以节省明年或未来的工作时间，就是一份有利息的工作。工作有利息，意味着你未来的工作时间会变少，多出来的时间，就可以去做别的事情，创造更多的价值。这提示我们：

- 不要轻易更换工作领域，否则以前积累的利息就作废了。只有在同一个工作领域，才可能产生长期积累，以前的工作为以后打基础，最终产生巨大的利息。
- 在职业生涯的早期，积累效果最好，最容易产生复利。开始积累越晚，产生复利就越少。
- 有些劳动没有积累效果，不会产生利息，比如重复性的机械劳动（快递、咖啡店员、门卫……），你明年还是要重复做这些事情。
- 最好的人生策略就是，找到你深感兴趣、可以长期做下去的领域，在上面投入大量的工作时间（包括质量和数量），然后随着年龄增长，享受以前工作的复利。

4、人工心脏

一个澳大利亚男子，植入了一颗人工心脏，已经活了100天，并且成功出院，创造了世界纪录。这相当于在胸腔植入一个血液泵，一天24小时推动血液循环。他是目前世界唯一一个带有人工心脏的人，也是世界第六例人工心脏植入。前五例的人工心脏都只是过渡，病人后来又移植了其他人的心脏。如果机器心脏以后技术成熟了，人类的寿命可望大幅延长。

5、如果没有人读博客，为什么要写呢？

让我们坦率一点吧，你写了一篇博客，根本就没有人读。至少，没有你想要的那么多读者。你把自己的想法倾注在文章，精心构思每个句子，选择合适的图片——然后什么反响也没有，没有点赞，没有分享，没有互动。

那么写博客的意义何在？首先，关于写博客，有两个误解。一个是只要我写出了好文章，读者自然就会来。不，他们不会来，网上有几十亿篇博客，好像浩浩荡荡的飓风一样，你的博客只是风里的一片叶子，谁会注意呢。另一个误解是如果没有人阅读，写作就是浪费时间。博客有自己隐藏的价值，你写博客不是为了别人的掌声，而是因为你自己的需要。

- 博客使人头脑清晰，它帮你理清思绪，锐化视角，当你写作时，你会思考得更好，当你思考得更好时，你会做出更好的成果。
- 博客的目标读者，其实不是互联网人群，而是未来的你，你的文章会让你看到自己思想的演变。
- 此外，未来也许有一天，某个真正需要你文章的人，会找到它。一篇有深度的文章比一篇病毒式传播的文章，影响力更持久。

写博客有点像街头摄影，你手拿相机，漫步在城市中，你看到一个场景——一个充满光、影、人性的瞬间，就拍下了它。没人关心你拍到了什么，但这不是你摄影的原因，你摄影是因为你看到了一些东西。写博客也一样。你写博客是因为你在思考，因为你在观察，因为你希望把它放在某个地方。如果有人读了，那就更好了，如果没有，工作还是完成了，这才是真正的重点。

6、红绿色盲

红绿色盲的患者，看不到红色和绿色。在他们眼里，这两种颜色都会变成黄色。大概每20个人里面，就有一个人有色盲或色弱问题。所以，设计界面的时候，使用红色或绿色必须非常谨慎，因为红绿色盲患者分不清。

下面的日历使用绿色和粉红色，表示特殊的日期。

但是，红绿色盲患者看到的是下面这样，根本分不清。

7、低代码编程受困于形式

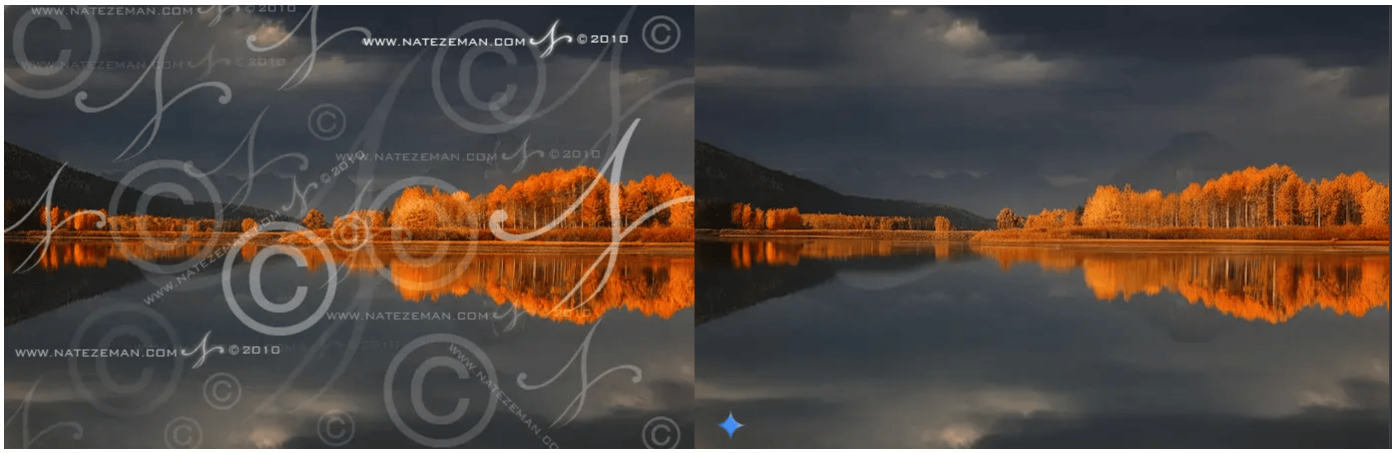
这十几年，一批批程序员前仆后继，去搞低代码编程（包括无代码编程）。这个想法很好，确实很多人需要，尤其不懂编程的人，这简直是生成程序的唯一可用方式。但是很奇怪，他们无一例外都失败了，开发出来的低代码工具，开始还有一些好奇的用户，很快就不来了，用户越来越少，后来即使开源了，也没人用。更奇怪的是，这似乎不是偶然现象，业界所有的低代码工具好像都不成功，没有哪一个受欢迎的应用程序是用低代码工具生成的。

优秀的作品都是形式（form）和功能（function）的统一。形式必须服从功能，功能决定了形式，英文叫做“form follows function”。对于优秀的程序员，只要弄清楚了底层，UI 就会显而易见。低代码编程的问题在于，它是先有 UI，再有代码。用户先拖拉生成 UI，系统再根据 UI 生成代码。这是本末倒置，让底层代码适配 UI，注定了两者都有问题：UI 是空想出来的，代码为了适配 UI，注定冗余和低效。

所以，优秀的软件不可能用这种方式生成，低代码编程不会成功。我认为，他说的很有道理。低代码编程解决不了这个根本缺陷，适用场景有限，大概只适合一些简单任务，或者生成原型，不会成为主流工具。低代码编程有先天缺陷，恐怕不会成功，程序员应该谨慎开发这类工具，付出的劳动很可能打水漂。

8、AI去除图像水印

很多美国用户在社交媒体上反映，谷歌新发布的 Gemini 2.0 Flash 模型，去除图片水印的效果极佳。虽然其他模型也能去除水印，但是 Gemini 2.0 Flash 似乎特别擅长这件事，而且它可以免费使用。



9、螺旋状的云

英国民众纷纷报告，夜空中发现螺旋状的云。英国气象局调查后宣布，那是猎鹰9号火箭发射时，快速旋转的箭体喷出的尾气。尾气在太空中瞬间冻结，经过太阳光反射，看上去像云一样。

八、闲情逸趣

梁文锋专题访谈

DeepSeek V3是来自杭州的量化基金公司幻方量化，一经发布，它就引起了国际范围的轰动。目前，它在大模型排行榜排名第7，在前十名里面，只有它是开源模型。它的训练成本很低，估计只有Meta的Llama 3.1 405B模型的1/11，而后者的效果还不如它。这也就是说，DeepSeek 找到了高效使用硬件、提高模型效果的方法。下面是幻方量化创始人梁文锋在专题访谈里说过的一些话：

- 我们要做的不是生成式 AI，而是通用人工智能 AGI。前者只是后者的必经之路，AGI 会在我们有生之年实现。

- 任何 AI 公司（短期内）都没有碾压对手的技术优势，因为有 OpenAI 指路，又都基于公开论文和代码，大厂和创业公司都会做出自己的大语言模型。
- 在颠覆性的技术面前，闭源形成的护城河是短暂的。即使 OpenAI 闭源，也无法阻止被别人赶超。我们把价值沉淀在团队上，我们的同事在这个过程中得到成长，积累很多know-how，形成可以创新的组织和文化，就是我们的护城河。
- 我们不会闭源。我们认为先有一个强大的技术生态更重要。
- 当前阶段是技术创新的爆发期，而不是应用的爆发期。大模型应用门槛会越来越低，创业公司在未来20年任何时候下场，也都有机会。
- 过去很多年，很多的中国公司习惯了别人做技术创新，拿过来做应用变现，自己等着摩尔定律从天而降，躺在家里18个月就会出来更好的硬件和软件。我们的出发点，就不是趁机赚一笔，而是走到技术的前沿，去推动整个生态发展。中国也要逐步成为贡献者，而不是一直搭便车。
- 大部分中国公司习惯 Follow，而不是创新。中国创新缺的不是资本，而是缺乏信心以及不知道怎么组织高密度的人才。我们没有海外回来的人，都是本土的。前50名顶尖人才可能不在中国，但也许我们能自己打造这样的人。
- 我们每个人对于卡和人的调动是不设上限的。如果有想法，每个人随时可以调用训练集群的卡无需审批。同时因为不存在层级和跨部门，也可以灵活调用所有人，只要对方也有兴趣。
- 我们选人的标准一直都是热爱和好奇心，所以很多人会有一些奇特的经历，很有意思。很多人对做研究的渴望，远超对钱的在意。
- 我们在做最难的事。对顶级人才吸引最大的，肯定是去解决世界上最难的问题。其实，顶尖人才在中国是被低估的。因为整个社会层面的硬核创新太少了，使得他们没有机会被识别出来。我们在做最难的事，对他们就是有吸引力的。
- 中国产业结构的调整，会更依赖硬核技术的创新。很多人发现过去赚快钱很可能来自时代运气，现在赚不到了，就会更愿意俯身去做真正的创新。
- 我是八十年代在广东一个五线城市长大的，我的父亲是小学老师，九十年代，广东赚钱机会很多，当时有不少家长觉得读书没用。但现在回去看，观念都变了，因为钱不好赚了，连开出租车的机会可能都没了。一代人的时间就变了，以后硬核创新会越来越多，因为整个社会群体需要被事实教育，当这个社会让硬核创新的人功成名就，群体性想法就会改变，我们只是还需要一堆事实和一个过程。

九、数字与言论

1、保护海底光缆不出事是不可能的，你唯一能做的就是建立大量冗余，在不同位置铺设数十根光缆可能比保护它们更便宜。——Hacker News

2、摩尔定律预测，芯片性能大约每年会翻一番。但是，AI 芯片的发展速度比这快得多。今天我们发布的 GB200 NVL72 芯片，运行 AI 推理的速度，比去年的上一代 H100 快了30倍，比10年前的芯片快了1000倍。我们正在超越摩尔定律，AI 适用超级摩尔定律。——黄仁勋

3、人的智力高低，未来不会像现在这样重要，AI 可以弥补人的智力。提出正确问题的能力，在未来比找到答案的能力更重要。——Sam Altman

4、我从未想到会被公司解雇，因为我的表现总是高于公司的期望。后来我明白了，在裁员期间，你是谁、你做什么似乎并不重要，在大多数情况下，裁员的决定是由那些不认识你的人做出的，对公司来说，我只是 Excel 表格中的一行。——《裁员改变了我》

5、我们越忙碌，就越能敏锐地感受到自己在生活，对生活也就越有想法。——康德，德国哲学家

6、程序员们不再互相提问，AI 回答了大部分问题。—— AI的数周相当于人类的几十年，自大模型问世后，StackOverflow日益冷清

7、有一句老话：创意很廉价，执行才是一切。然而，AI 颠覆了这个说法，执行现在很廉价，整个开发时间和交付速度的概念都不同了。未来属于那些有想法、还能动手去做的人。——Geoffrey Huntley

8、在研究和学术领域，成功往往不属于最先理解的人，而属于理解得最好的人，真正的优势来自于深刻、基础性的见解。——《我在麻省理工学院的时光》

9、以前的球票、音乐会票、景点票、电影票都是纸质的，现在全改成数字的。我们的过去都保存在手机里，再也没有纪念物了。——彭博社

10、我见过的最好的工程师，是那些愿意在周末花几个小时构建一个现有软件的自己版本的人。这就是你获得创新和进步的方式。如果你不了解系统的工作原理，就无法找到改进的地方。——《AI 让开发者变蠢》

11、使用 GitHub Copilot 后，我得了一种叫做“Copilot 延迟”的病。这种病指的是工程师在每次操作后都会暂停，等待 AI 提示他们下一步该做什么。很多工程师有了 AI 以后，就做不到只靠自己了，要靠 AI 告诉他们下一步。这类似于初级程序员在刚开始时，依靠资深的同事的指导开展工作。—— 《AI 让开发者变蠢》

12、如果你成功了，记住你要去哪里，记住你来自哪里，并选择你要成为什么样的人。——《五周的独自创业》

13、开源运动的人们完成了不可能的任务。他们创造了整个百科全书、地球上最成功、使用最广泛的操作系统、软件库和无数应用程序。他们对公共资源的贡献甚至在科幻小说中都难以想象，其中一些系统应该被视为世界的数字奇迹。—— 《自由软件为了谁？》

14、中国睡眠研究会统计，2025年中国人夜间平均睡眠6.85个小时，比去年增加6分钟。入睡时间平均为0点18分，比去年晚了17分钟。—— 新浪新闻