

NaniAI 安装与使用指南

一. 下载步骤:

官网(<https://ailrxi.com/products/nani-ai.html>)下载软件, 下载后如果浏览器弹出未知应用是否保留, 请选择**保留**; 部分浏览器可能会误屏蔽, 请在屏蔽警告的删除下拉**选择保留**。

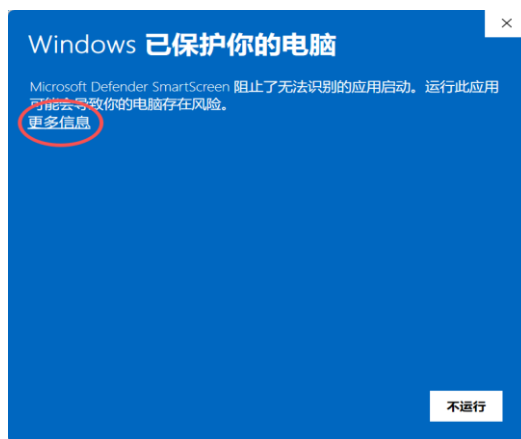
二. 安装步骤:

Step1: 解压 NaniAI.zip 压缩文件, 内有:

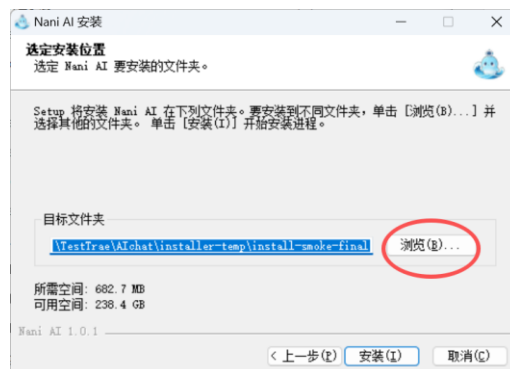
- ①安装向导文件: Nani AI Setup xxx(版本号).exe
- ②NaniAI 首次安装与使用指南.pdf (本文件)

Step2: 双击安装向导文件.,例如 Nani AI Setup 1.0.1.exe

由于部分 windows 验签严格时候可能会出现 SmartScreen 阻止, 点击"更多信息",弹出下方 "仍要运行" 按钮, 点击 "仍要运行" 即可。



Step3: 安装路径可以自定义变化,也可以默认,一路点击下一步即可完成.



关于版本更新说明: 当官网发布新版本, 有效订阅用户软件账户面板将会检测新版本信息, 可在软件点击“立即更新”, 后台会自行下载新版本, 然后安装覆盖旧版本; 下载期间不影响用户软件使用, 下载完成后会提示立即更新还是稍后更新, 请用户选择; 选择稍后更新将本地保存新版安装包, 等待下次用户自行点击更新。由于服务器宽度有限, 用户会随机等候下载, 如不想等待, 可点击“查看更新”可跳转官网, 从官网下载, 如遇到网速限制, 建议从官网“备用链路”下载。

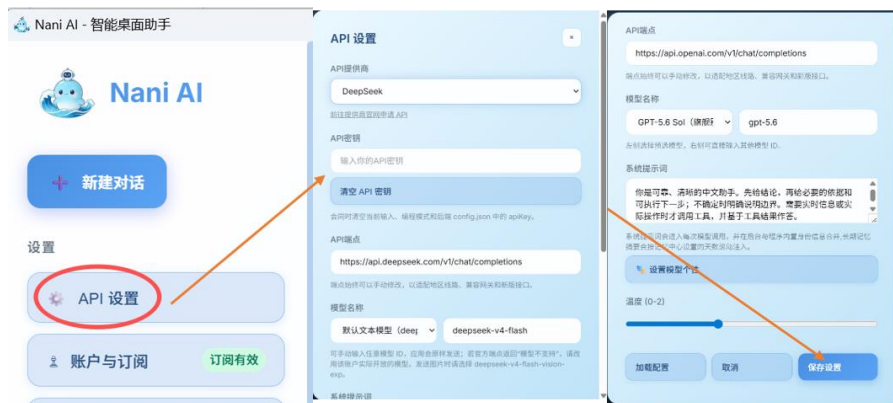
三. 快速使用指南:

本应用基础功能免费，但是部分高级功能需要开通订阅，包括：聊天的多步骤、MCP 设置、技能管理、定时任务，编程模式，桌面智能宠物功能。

- 布局: 聊天-编程-桌宠模式切换（鼠标点击切换）、个性化颜色主题选择，打开/关闭左侧设置面板、导出聊天记录、打开内置浏览器窗口或文本阅读窗口；



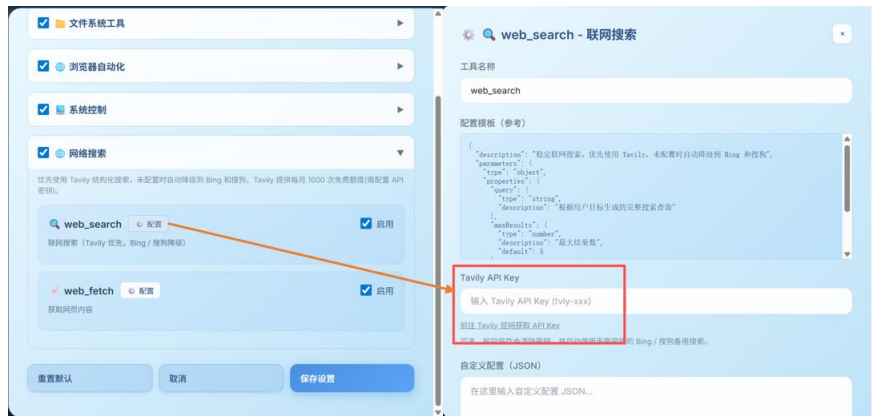
- **API 设置:** 使用前须先设置大模型 API 接口，软件内置主流厂商 API 接口和模型格式,也可自定义 Open AI 格式的 API 接口，兼容本地部署的大模型 Ollama 接口；(通过点击“前往提供商官网申请 API”,注册获取 API 后,选择或填入对应大模型接口信息和模型名称,填入 API key), 点击保存; 此时也可同步修改系统个性化设置,增强个性化记忆与行为(也可保留默认设置). **p.s:** 国内外有部分免费提供大模型 API 的渠道，请自行了解，也可选择本地部署大模型（取决于本地机器条件，各主流厂家 API 注册地址和 Ollama 部署本地模型的简易教程见本说明后附录），但模型质量对智能体应用体感比较明显。欢迎有算力的资源方可联系合作，以提供最便捷极致和便捷的用户体验。



- **账户订阅:** 软件基础功能免费，部分高级功能需要订阅，订阅范围包括: MCP 设置，技能管理，定时任务，编程功能页，桌宠功能页。订阅可从面板左侧“账户与订阅”打开登录，已注册的用户可采用注册的账户和密码在软件登录，未注册用户，点击“注册账户”跳转注册页面，并使用邮箱激活；订阅/续费可从网页登录后的账户中心发起，也可从软件账户面板灯“订阅/续费”按钮跳转发起，订阅完如果没有及时自动刷新订阅状态和日期，可手动点击“刷新状态”。



- **内置插件:** 软件已经内置了文件/文件夹读写, 浏览器自动化, 电脑本机系统命令执行和网络搜索功能, 建议为了增强网络搜索稳定性, 可以注册 Tavily 账号 (<https://www.tavily.com/>), 获得搜索的 API Key, 填入此处后保存, 每人每月有 1000 次免费网络搜索额度。



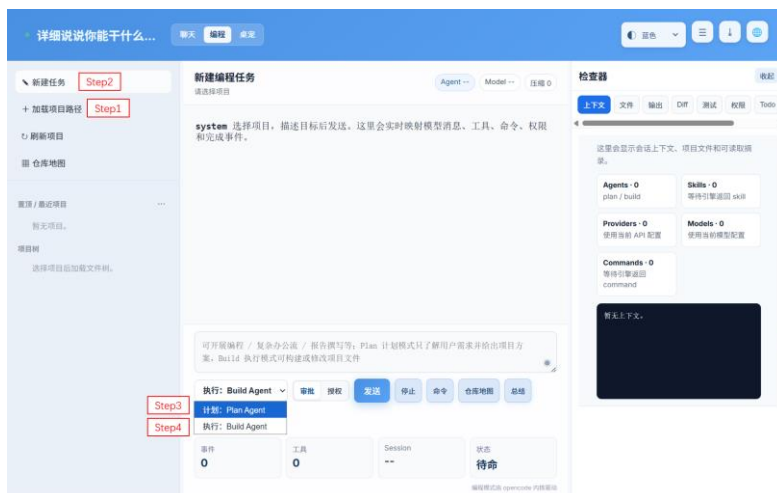
- **模式:** 本应用分为聊天\执行模式, 编程模式和桌宠模式,各自定位如下:
 - (a) **聊天\执行模式:** 日常聊天办公, 轻度浏览器控制和本地执行等, 主要特点是响应快, 但是运行的任务循环上限比较低, 以节省 token。如果涉及多步骤复杂问题, 或普通模式效果不好, 建议勾选“多步骤”, 回在执行器充分思考规划步骤, 并且执行循环检测任务完成, 但是效率比较慢; 需要知识库知识, 可以勾选知识库, 但是本知识库是简易关键词检索匹配, 属于轻量型。



- (b) **编程模式:** 编程模式驱动具有较强的智能体执行能力, 可以进行专业的程序修改和新建程序项目, 也可以用来执行办公文件撰写, 报告整理撰写, 数据整理等。第一步加载已经存在或新建的项目路径, 第二步新建任务, 如果任务比较复杂或者需求不清晰, 建议可以选择“计划: Plan Agent”模式, 与 AI 聊清楚需求, 并且让 AI 给出具体方案, 然后下拉选择“执行: Build Agent”模式, 直接告诉 AI, 让其根据已给的上述方案执行; 如果需求比较明确或简单, 可以直接选择“执行:

Build Agent”模式。

编程模式左侧是文件树和文件管理区域，右侧面板是编程 AI 的辅助详细信息流、设置和交互选择区。



- (c) **桌宠模式:** 桌宠模式，将召唤我们可爱的桌面宠物 Nani 海宝，内置形象还有开放资源的第三方动态猫形象和动态企鹅形象。可以在这里设置我们的桌宠行为特性，甚至 DIY 我们自己的自定义形象，点击“显示桌宠”，萌萌哒桌宠就会在我们电脑桌面显示并到处游走，甚至与你互动。同时桌宠兼具 AI 聊天和执行功能。来开启你的桌面宠物吧！



■ 其他设置:

(1) MCP\SKILL\定时:

软件左侧面板有一些预置 MCP 模板、skill 模板，可以进行加载，其中 MCP 支持 npx, uvx 和 https 模式的 MCP 安装指令 (npx, uvx 的依赖请自行安装)，也可以通过内置 MCP 搜索引擎搜索需要的 MCP；可以通过面板的技能管理，通过程序内置的 skill 搜索引擎搜索需要的技能并通过“导入”按键，软件将自动安装；也可以通过聊天，让 AI 帮你搜索和安装 MCP 和 skill（多步骤模式执行成功率较高）。

定时任务：将设置时间，让 AI 提醒或自动执行任务；

(2) 记忆与知识管理:

软件有轻量的关键词检索知识库，可以导入 PDF, txt 等文件，聊天时候勾选知识库可以在 AI 上下文结合；

记忆中心：可以设置个人关键信息，让 AI 了解你，同时可以滚动保持聊天摘要，让 AI

越用越懂你;

(3) 多模态: 软件当前版本考虑安装文件大小, 没有内置 OCR 视觉识别, 如果需要可通过 skill 自行安装, 软件可以发送 pdf 等文件, 会内置转化处理, 图片需要对接的 API 为多模态大模型, 默认对接的格式以 deepseek-v4-flash-vision-exp 的 Base64 编码格式。

■ 帮助与支持:

通过软件内置的帮助中心进一步了解软件使用方法和功能;

或通过我们官网的教程中心获得相关教程视频等: <https://ailrxi.com/training/index.html>

四、API 信息:

软件目前不提供大模型算力, 用户可根据自己需求和实际情况, 选择接入第三方厂商或平台的大模型算力 API 或本地部署算力, 软件已经内置了多种主流大模型厂家 API 格式, 用户也可以自定义, 默认支持 OpenAI 格式的 API。

■ 国内常用 API 信息

DeepSeek: [DeepSeek 开放平台](#)

阿里千问: <https://help.aliyun.com/zh/model-studio/get-api-key>

智谱: [智谱 | BigModel 平台](#)

Kimi: [登录](#)

MINIMAX: [MiniMax AI](#)

■ 免费 API 信息

官网提供了大部分国内外免费的大模型 API 的渠道, 点击: [免费开源 AI API 聚合目录](#)



■ 本地部署简易指南:

登录 Ollama 官网了解相关详情: [Ollama](#)

Step1: 如果没有安装 ollama 和模型, 需要 powershell 执行安装 ollama:

```
irm https://ollama.com/install.ps1 | iex
```

或者官网点击直接下载对应版本;

Step2: 拉取大模型部署本地, 在本地 powershell 运行: ollama run <模型名>

本地电脑具体能安装什么模型或自己想要什么模型, 可以从 ollama 官网的 model 库搜索查看, 点击某模型后进去可以复制模型名称, 带-cloud 的模型为 ollama 提供算力的云模型, 不占用本地电脑算力和磁盘, 但是每周免费使用量有限制, 数据会上网。不带-cloud 的

模型为下载本地运行的模型，文件比较大，本地电脑性能能否驱动，建议查看自己电脑配置属性或让 AI 帮忙分析。两种模型都是相同的 ollama api 格式，只是模型名称不同，本软件已经内置支持 ollama 格式的 API。

Step3: 下次启动 ollama 和大模型：命令：ollama serve，或点击开启 ollama 的 exe
列出本地已安装模型命令：ollama list 运行大模型命令：ollama run <模型名>

Download Ollama


macOS


Linux


Windows

```
irm https://ollama.com/install.ps1 | iex
```

paste this in PowerShell

or

Download for Windows

Requires Windows 10 or later

Models Docs Pricing

Search models

Cloud Embedding Vision Tools Thinking Popular

glm-5.3
Z.ai's flagship model and the most capable open-weights model for coding, with major gains on long-horizon agentic tasks.
[tools](#) [thinking](#) [cloud](#)
34.5K Pulls 1 Tag Updated 1 week ago

glm-5.3-flash
Z.ai's first natively multimodal model, approaching Claude Opus 4.8 on coding and agentic benchmarks with just 18B active parameters.
[vision](#) [tools](#) [thinking](#) [cloud](#)
85.1K Pulls 1 Tag Updated 1 week ago

deepseek-v4-flash
DeepSeek-V4-Flash is a preview of the DeepSeek-V4 series, a Mixture-of-Experts model with 284B total parameters and 13B activated, built for efficient reasoning across a 1M-token context window.
[tools](#) [thinking](#) [cloud](#)
427K Pulls 2 Tags Updated 1 month ago