

AI 智能体

从大模型到你的个人助理



AI 智能体讲座

01 **大模型与智能体原理**
它是怎么工作的，为什么会出错

02 **主流大模型、厂商与智能体产品**
谁是谁，谁适合干什么，大概多少钱

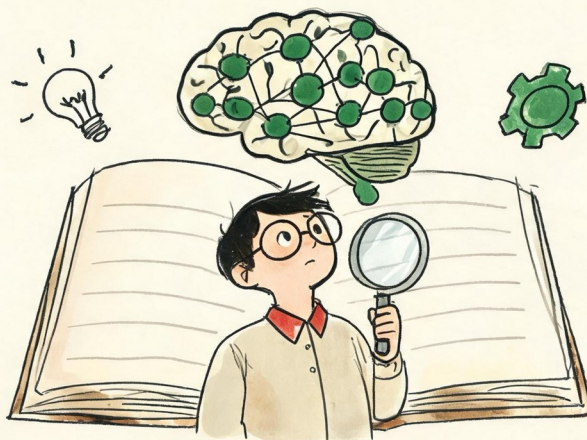
03 **以小荷为依托：安装、配置与使用**
把这条路走通——数据在自己手里



PART 01

大模型与智能体原理

第一部分 原理



先建立直觉，再谈能力与边界

三分钟看懂：AI、机器学习、深度学习

原理

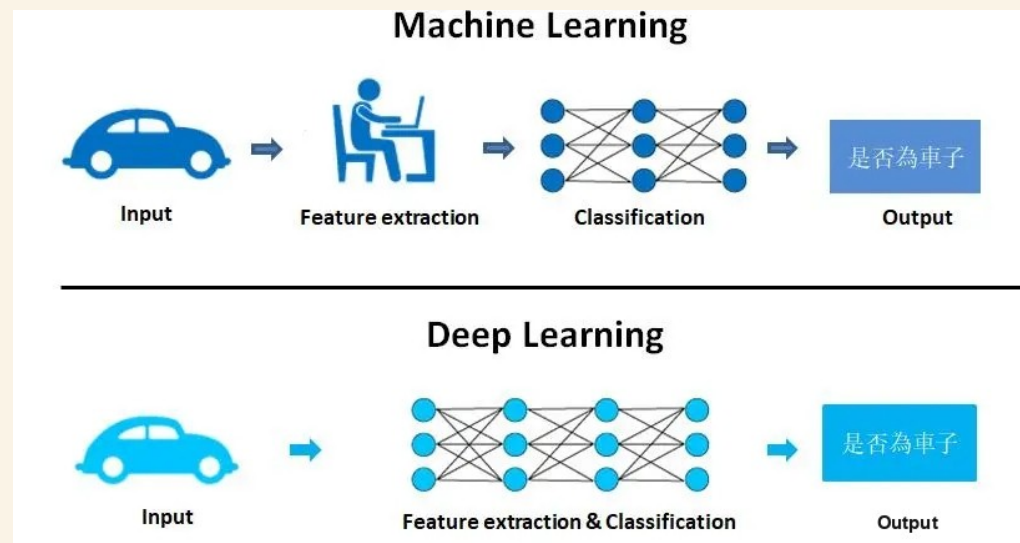
人工智能 让机器做需要智能的事，最外层的筐

机器学习 不写死规则，用数据让它自己找规律

深度学习 用多层神经网络做机器学习

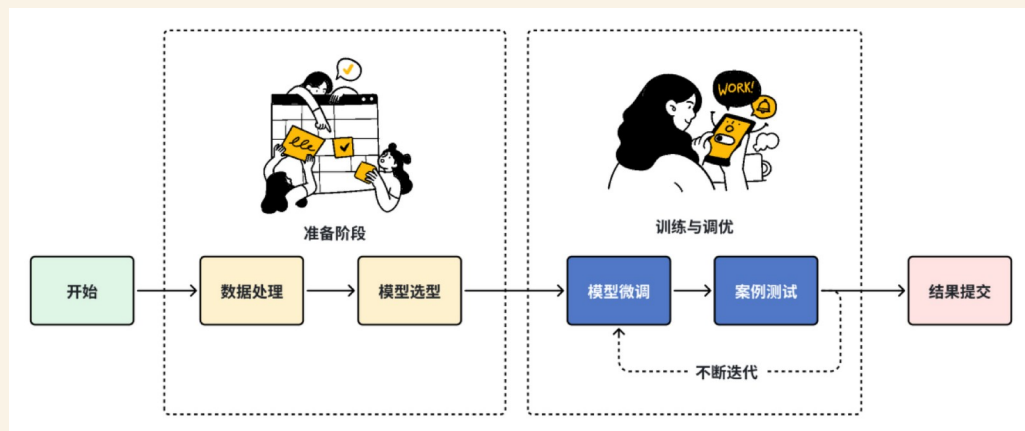
大模型 深度学习里**参数特别多**的那一类

层层包含：越往下越具体



从熟悉的函数说起：输入决定输出

原理



- 工程里我们习惯：给定输入，按公式得到输出
- 神经网络的骨架**也是一个函数**，只是参数多到写不出来
- 公式是我们推导的，参数是它自己学出来的

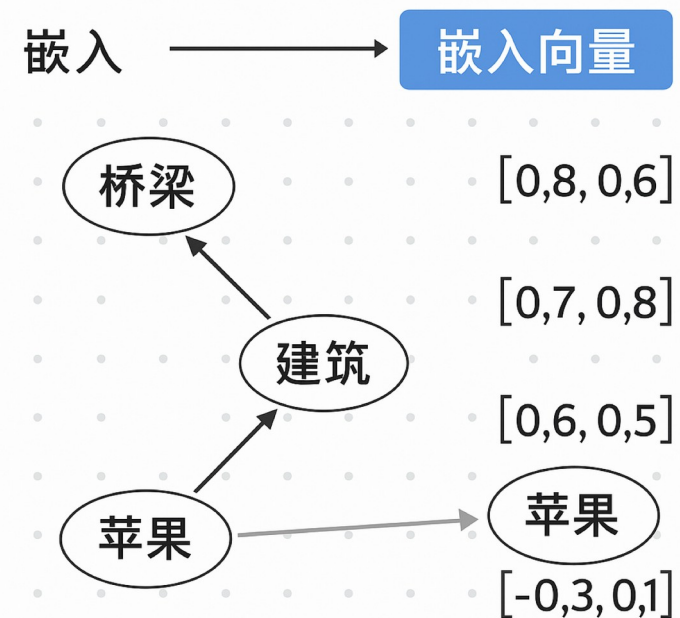
以前把物理规律写成公式，现在把规律交给数据去拟合

神经网络：一堆带权的连接

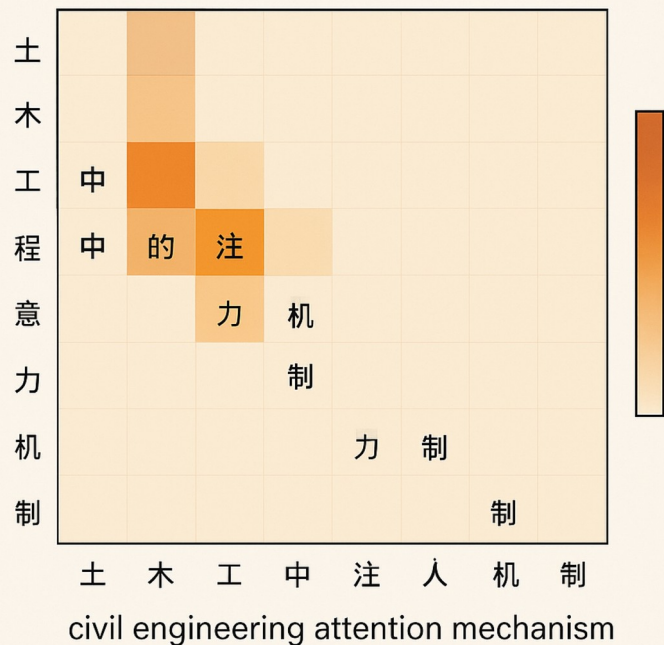
原理

- 基本单元：输入乘权重、求和、过一道非线性函数
- “训练”就是反复调整这些权重
- 它不存储知识条目，**知识藏在权重的数值里**

把一件事表示成一组数，模型比的就是这组数之间的距离



土木工程中的注意力机制



- 自注意力：每个词按“相关程度”给全句其他词加权
- 长距离依赖不再是难题，且天然适合并行计算
- 生成是逐词的：每次只预测“**下一个最可能的词**”

左图是同一套机制用在工程文本上：颜色越深，越是它重点参考的地方

大模型的三段式训练

原理

01 预训练

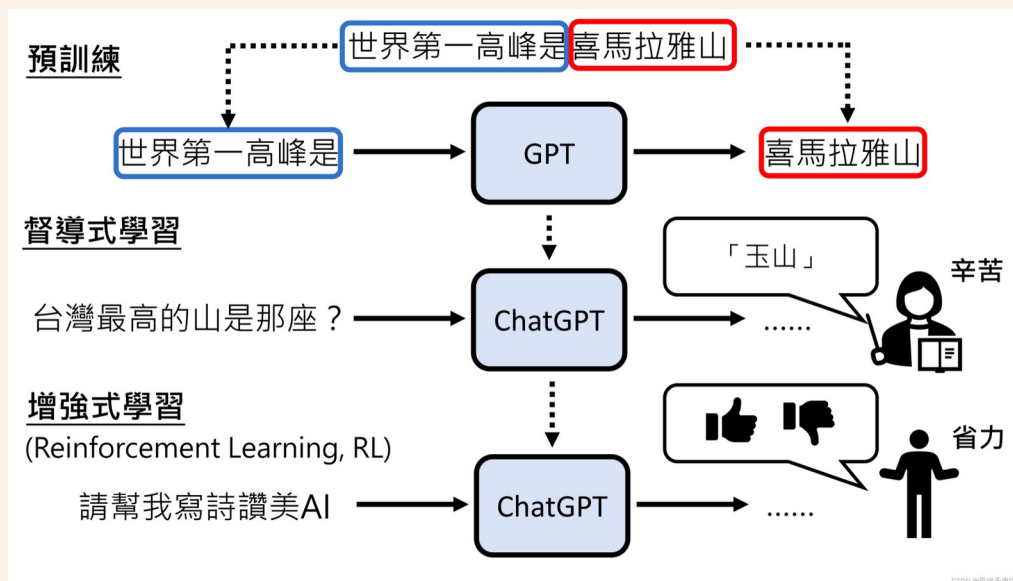
读海量文本，学语言规律与世界常识

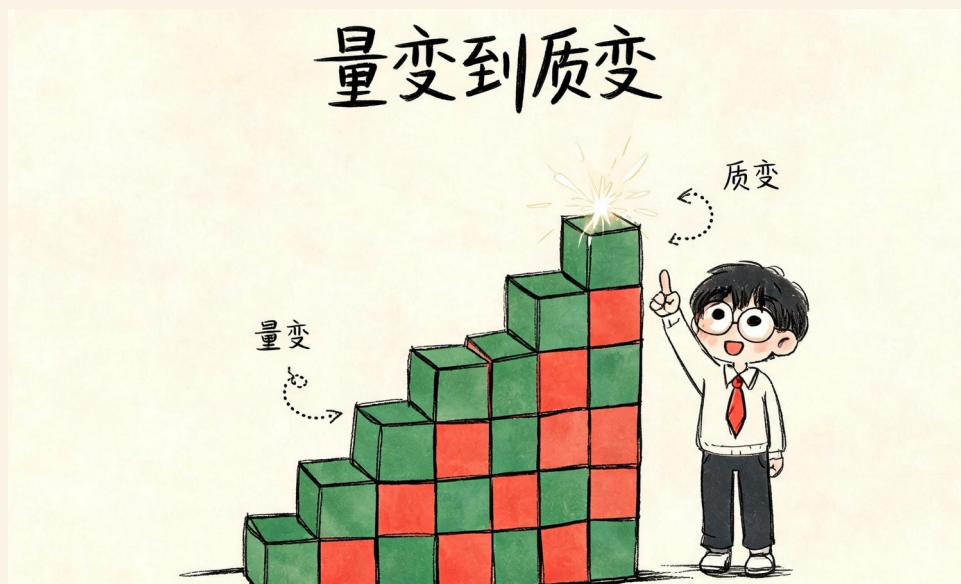
02 指令微调

学会“听懂要求并按格式回答”

03 人类反馈强化

校准偏好、礼貌与安全边界





- 参数、数据、算力同步放大，能力不是线性增长
- 规模跨过某些门槛，会出现训练时**没教过**的新能力
- 这让“堆规模”在一段时间内成了最有效的策略

代价同样真实：训练成本、推理成本与能耗，都随规模一起涨



- 它输出的是“最像样的下一个词”，不是查证过的事实
- 概率上通顺，**不等于**事实上为真
- 越冷门、越具体的事实，越容易编
- 参考文献、法条、数据，是最需要人工核实的重灾区



- Token 是模型处理文本的最小单位，中文大致一个字到一个词
- 上下文窗口 = 一次能“看见”的 token 上限
- 主流旗舰已到**百万级**，约合数十万字
- 窗口不等于记忆：超出就被丢弃

读进去的、写出来的，都按 token 数算钱

① 提示工程 Prompt Engineering

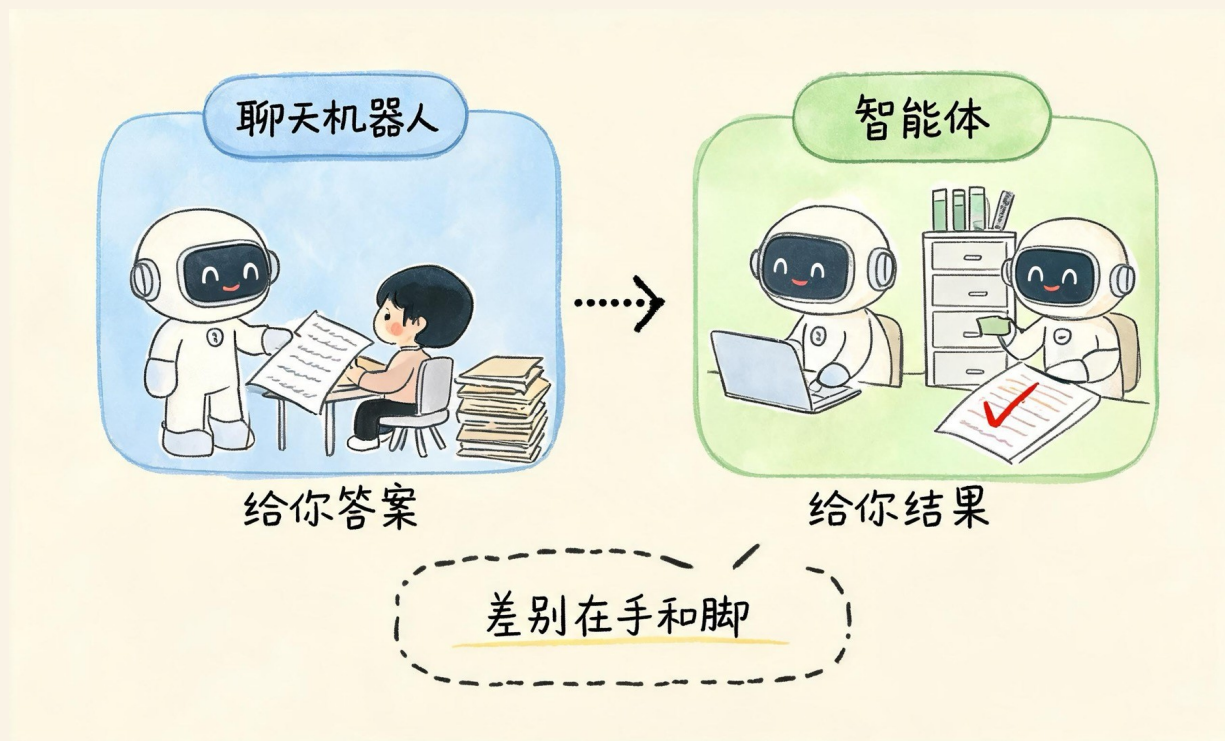
当场把要求说清楚——给角色、给约束、给例子（few-shot）、要求分步（思维链）。最快，不用改模型

② 检索增强 RAG

把资料库挂给它，回答前先翻文档——不放进上下文，**只取相关内容**，装得下、查得准

③ 微调 Fine-tuning

拿你的数据改权重，一劳永逸；但最贵、最慢，还要维护

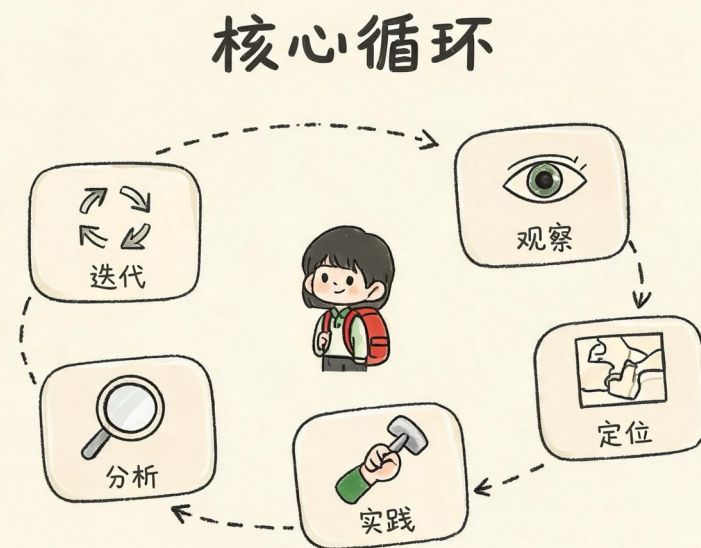


- 聊天机器人给你答案，智能体给你结果
- 差别不在模型，而在外面那圈“**手脚与循环**”
- 智能体 = 大模型 + 规划 + 工具 + 记忆 + 循环
- 同一个模型，装上手脚之后能干的事量级不同

核心循环：感知—规划—行动—观察—反思

原理

- | | | |
|----|----|---------------|
| 01 | 感知 | 读入任务与当前状态 |
| 02 | 规划 | 拆成步骤，决定下一步 |
| 03 | 行动 | 调用工具去执行 |
| 04 | 观察 | 拿回结果，判断是否符合预期 |
| 05 | 反思 | 不达标就调整做法，回到规划 |

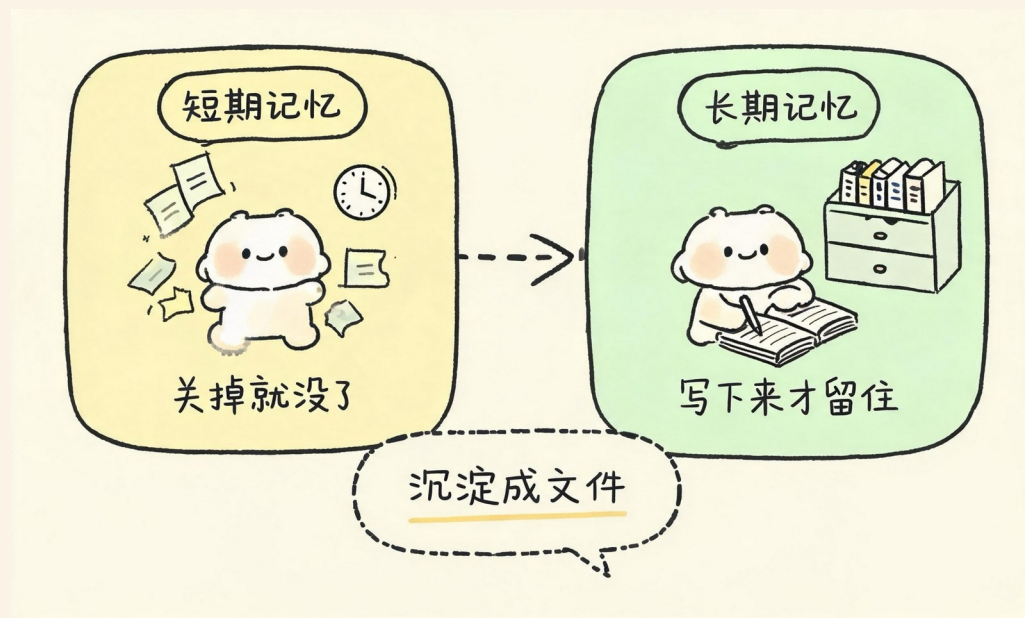


工具调用：让模型的手伸出屏幕

原理



- 模型不直接做事，它输出一个结构化的“调用请求”
- 宿主程序执行，把结果回传，模型接着往下走
- 读文件、跑代码、查网页、发邮件，本质都是工具
- 权限边界设在这里：**能碰什么，由你决定**



- 短期记忆：当前对话窗口里的内容，关掉就没了
- 长期记忆：写进外部文件，跨会话留存
- 好的助理会把结论**沉淀成文件**，而不是每次重问
- 记忆也带来风险：写进去的敏感信息会被反复读到

拆成小步



做个讲座



找素材



定大纲



出页面

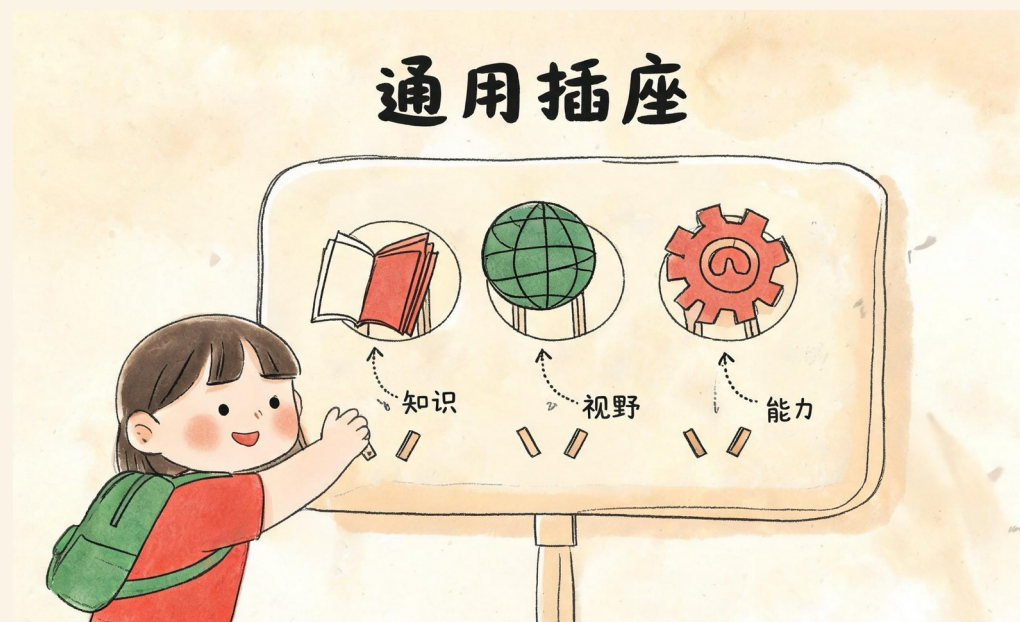


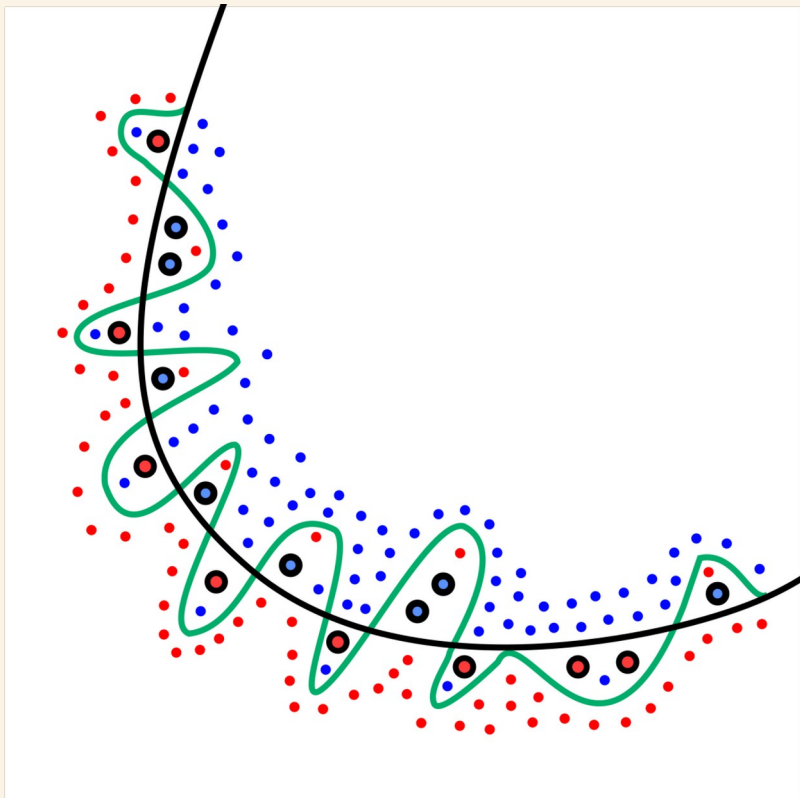
校对

- 每一步都要**可验证**，否则错了也发现不了
- 长任务要能断点续做
- 人留在关键节点：定方向、做取舍、把关

- 没有协议时，每接一个系统都要单独写适配
- MCP 把“模型怎么连外部能力”标准化
- 一次接好，多个模型与客户端都能复用
- 与技能分工：一个管**连得上**，一个管**怎么用**

像 USB-C：插口统一了，设备换谁都能接上





- **分工**：不同角色各管一段，并行推进
- **互查**：一个干活、一个挑错，可靠性提升
- 代价是协调开销：每个代理都要重新理解上下文

简单任务不要上多代理——成本翻倍，收益有限

这一节的边界：四个必须记住的风险

原理

安全

给了工具权限，就可能越权操作

隐私

数据发到公网模型，等于出境

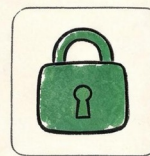
成本

按 token 计费，长任务容易失控

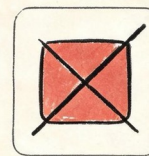
可靠

幻觉进了正式产出，责任在你不在模型

四个风险



账号风险



信息泄露



财产损失

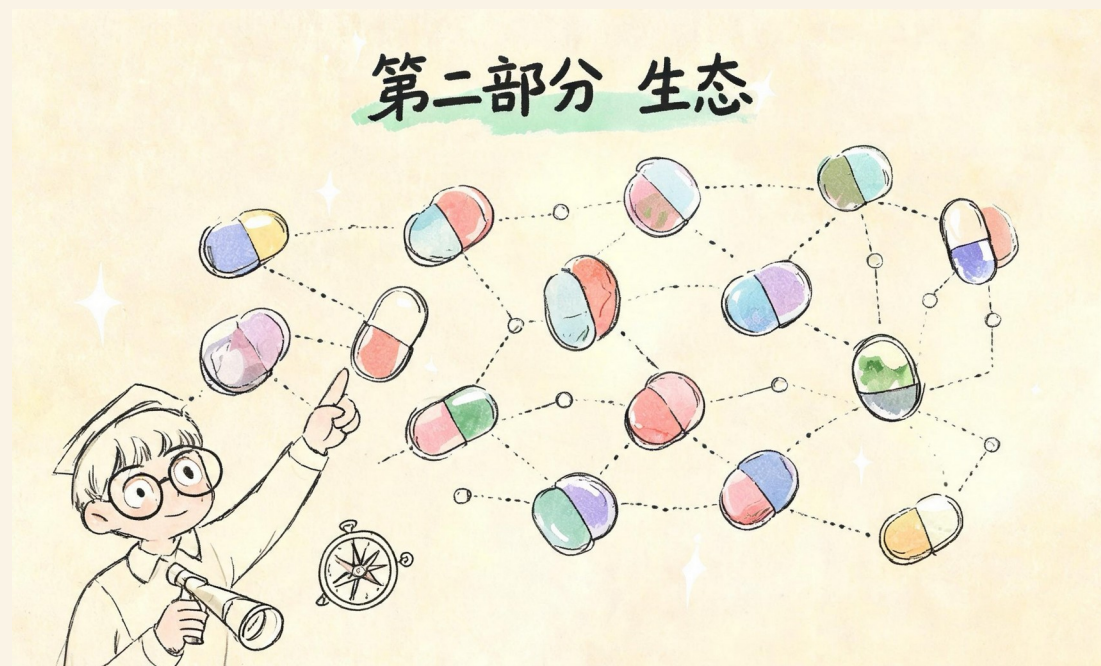


未知风险



PART 02

主流大模型、厂商与智能体产品



数据截至 2026 年 10 月，以各厂商官网为准

一张图看懂当下的模型版图

生态

海外第一梯队

Anthropic

OpenAI

Google

国产主力

DeepSeek

通义千问

Kimi

智谱 GLM

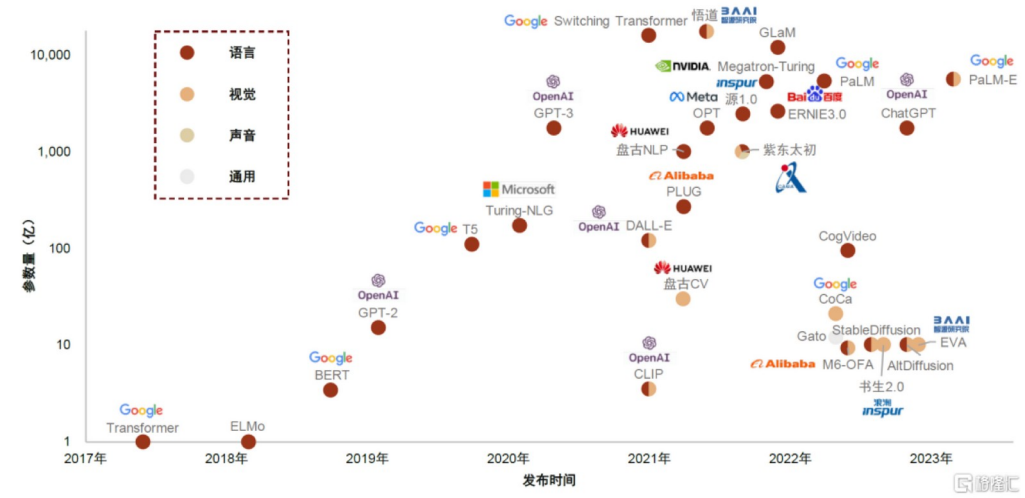
豆包

闭源 API 路线

能力更强，免运维，按量付费

开源权重路线

可私有部署，数据不出内网



最强推理

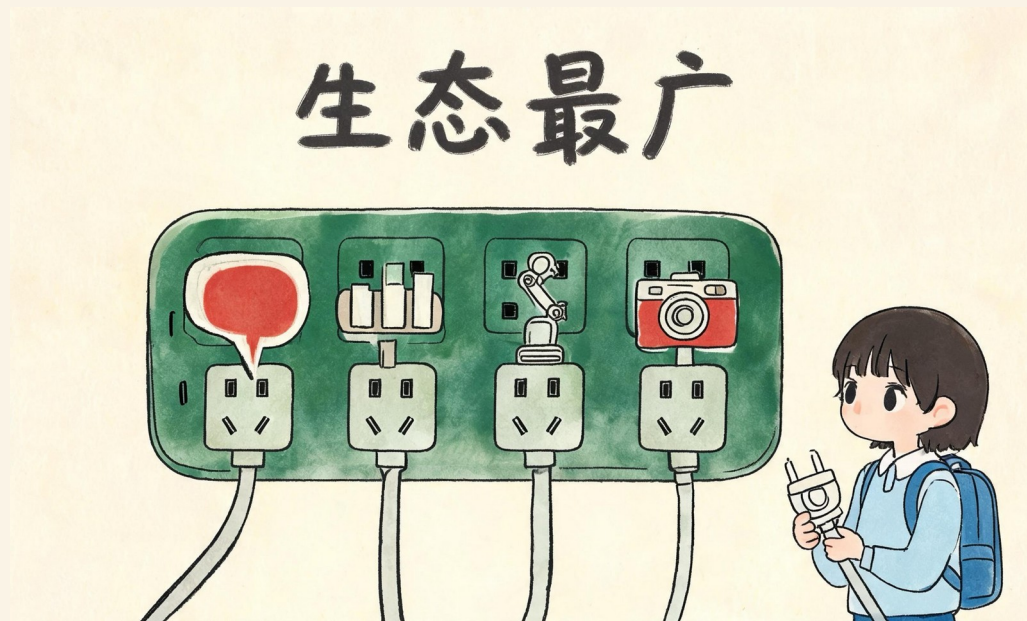


Claude Fable 5	最强，长程智能体与复杂推理	\$10 / \$50
Claude Opus 5	复杂智能体与编码主力	\$5 / \$25
Claude Sonnet 5	速度与智能的平衡点	\$3 / \$15
Claude Haiku 4.5	最快最省，适合简单任务	\$1 / \$5

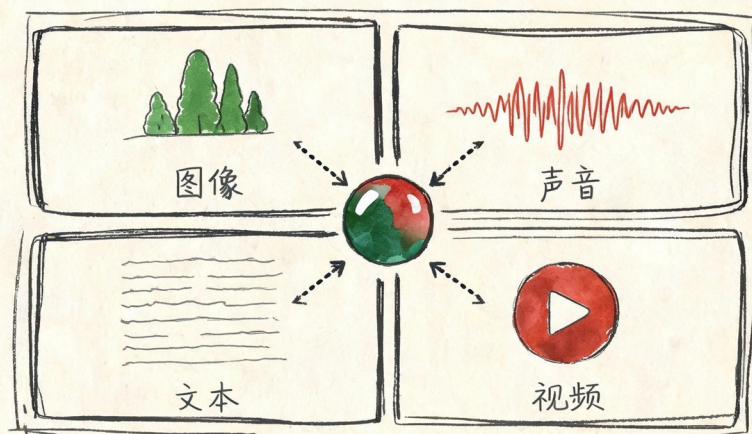
价格为每百万 token 的输入 / 输出（美元）

- 版本迭代最快，旗舰与轻量版并行
- 上下文普遍达到**百万 token** 级
- 优势在生态与应用面：插件、计算机操作、企业集成
- 价格约 **\$2.5–5 / \$15–30**，另有更低的 mini 档

选它多半是因为“什么都能接上”



原生多模态



- Pro 与 Flash 两档：一个拼能力，一个拼性价比
- 原生多模态：文本、图像、音频、视频统一处理
- 上下文 100 万 token 级，深度绑定 Google 生态
- Flash 档定价激进，约 **\$1.5-2 / \$9-12**

处理图片、表格、视频这类工程材料，它是最顺手的入口之一

海外三强：怎么挑

生态

要最强的推理与长程智能体

Claude 旗舰

长任务、多步工具调用、复杂代码改动

要生态最广、开箱即用

GPT 系列

第三方工具与平台适配最齐全

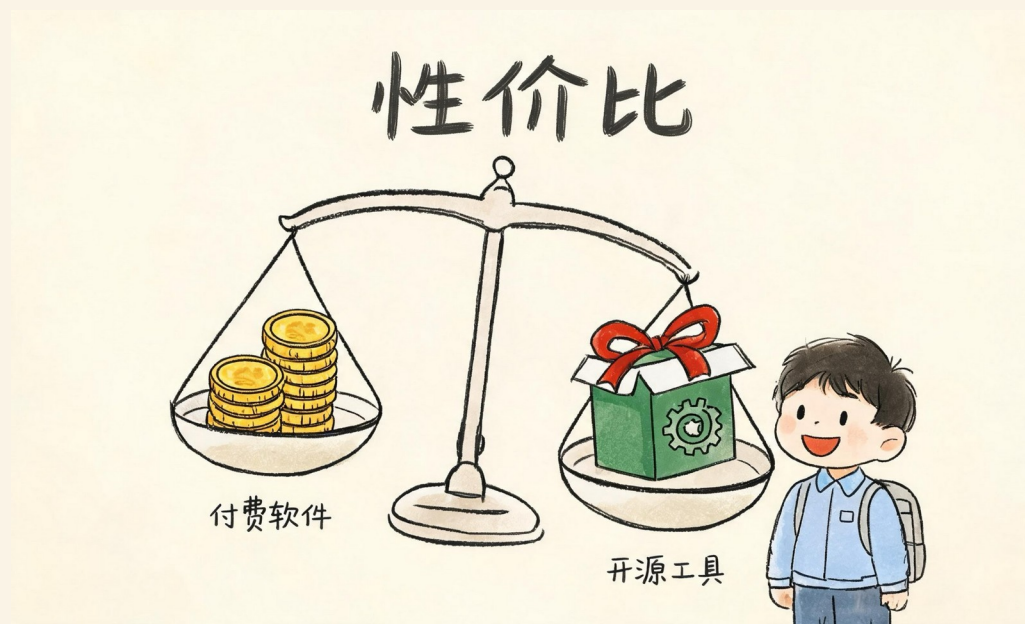
要多模态与性价比

Gemini Flash

图片、表格、视频处理，单价低

三家都支持批处理约五折、缓存读取约一折





- V4 系列分 Pro 与 Flash 两档，均支持百万级上下文
- 推理与代码能力突出，价格**显著低于海外旗舰**
- 开源权重，可私有化部署
- 代价：高峰时段算力紧张，响应可能变慢

同样任务，成本常常只有海外旗舰的几分之一

- 从 Max 到**可本地跑的小尺寸**，型号覆盖最完整
- 数学推理全球前列，文档解析与 PPT 生成好用
- 深度打通阿里云，企业落地路径成熟
- 开源版本下载量与衍生模型数量领先

一台带显卡的机器就能跑它的中小尺寸版本





- 招牌能力是**超长文档解析**与研究报告生成
- 上下文覆盖数十万到百万 token
- 定价处于国产高端档，适合质量优先的长文档任务
- K 系列开源权重，兼顾自部署可能

把一整套规范、几份长报告一起丢进去做交叉比对，是它的主场

- 逻辑推理与代码表现突出，开源友好
- 长程任务自主闭环能力强，**政企合规性最好**
- 开源权重 + 订阅双路线，私有化落地案例多
- 上下文 20 万 token 级，日常任务足够

对合规要求高的单位，它的私有化方案常被优先考虑





豆包

多模态与 C 端交互最流畅，绑定抖音与剪映生态；API 定价处于中低档

MiniMax

走**极致低成本**路线，适合量大、不要求顶配的场景

共同点：把“够用”做到了很低的价格

国产横向：价格已经明显分层

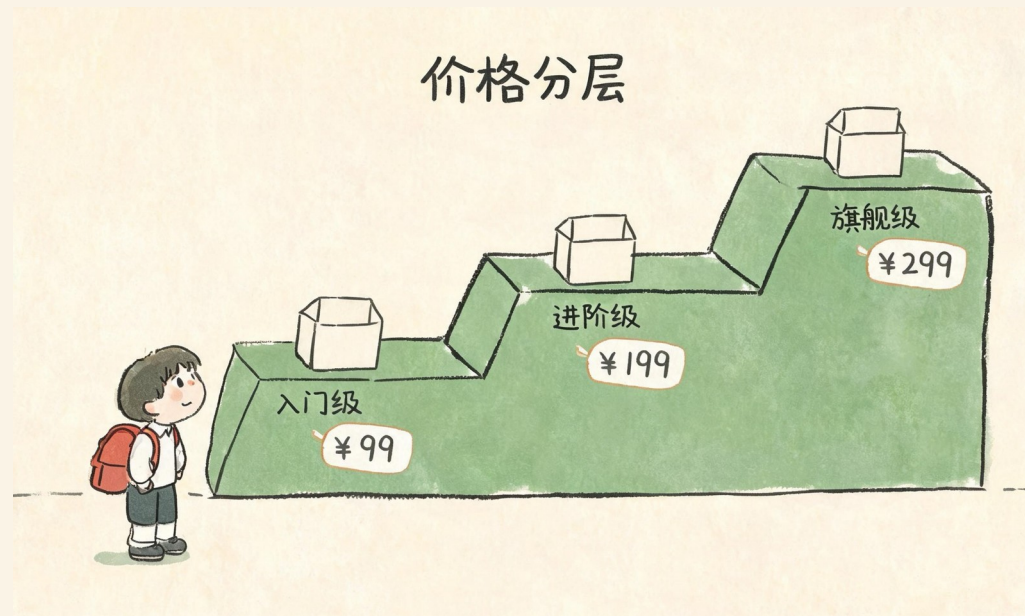
生态

极致性价比 DeepSeek、MiniMax

中低价位 通义千问、智谱 GLM、豆包

高端档 Kimi 长上下文系列

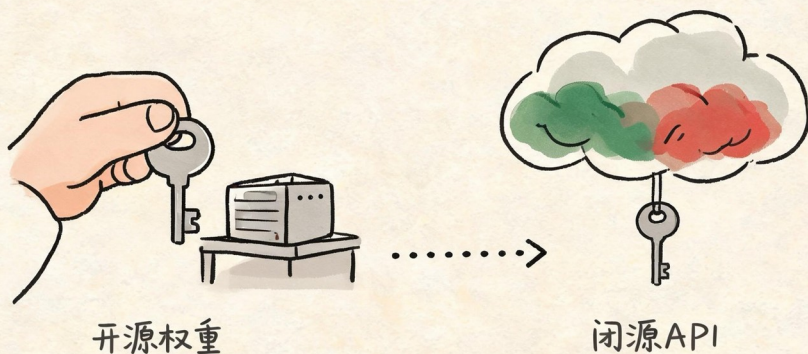
同一档内选谁，看生态与任务匹配，不看参数大小



开源还是闭源：权重在手意味着什么

生态

钥匙在哪



开源权重

可私有部署、可微调、数据不出内网、长期可控。运维与硬件自己扛

闭源 API

能力通常更强、免运维、按量付费、迭代最快。数据要出内网

常见折中：敏感数据用本地模型，写作分析用云上旗舰

两条新赛道：多模态与推理模型

生态

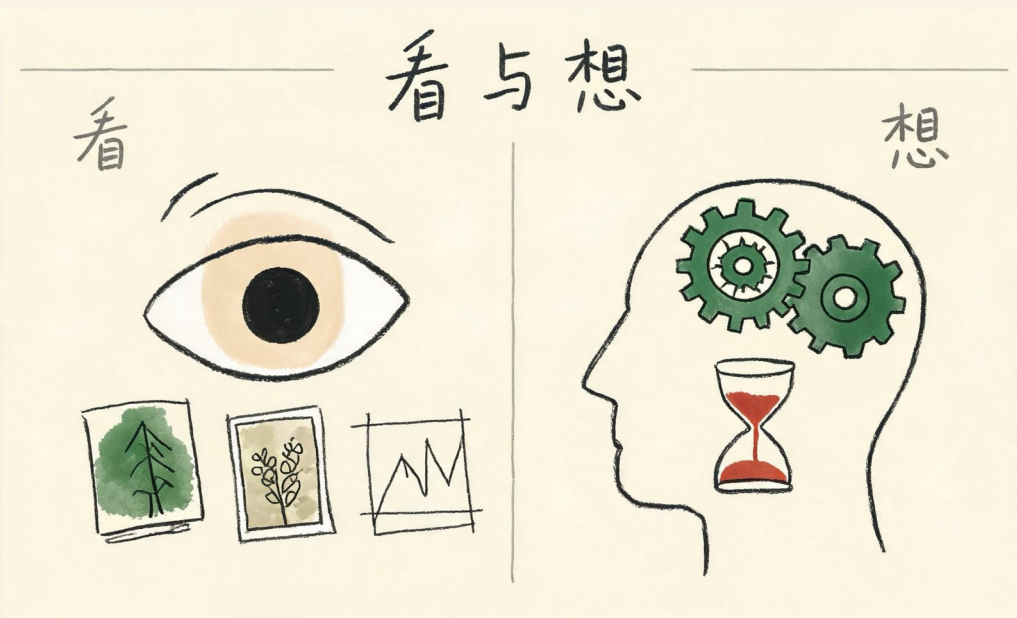
多模态

能看图、看表、看视频——图纸、实验数据、现场照片的入口

推理模型

先想再答，数学与复杂推理显著变强；代价是**更慢、更贵**

任务难度决定用哪一档，别一律上最强





01 数据能不能出内网？

不能 → 选开源权重，本地部署

02 任务要多深的推理？

难 → 旗舰模型；日常 → 小模型 + 好提示

03 长文档，还是多轮工具调用？

长文档看上下文长度，工具看代理能力

04 预算多少？

成本敏感 → 国产 API；质量优先 → 海外旗舰

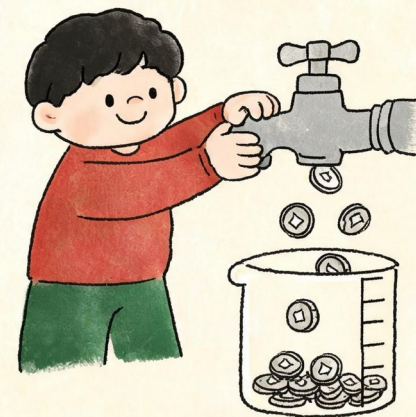
- 输入输出分开计价，**输出贵得多**；长对话每轮都重发历史，成本滚雪球

这份 PPT，全程 ~2 万 token

大纲 + 63 页文字 + 给 AI 的指令 ≈ **2 万 token**；用 Claude Sonnet 约 **1 毛钱**——一次讲座，几分钱

再给个尺度：审一篇论文 ~1 万 token（几分钱）；
读一本书 ~20 万 token（几块钱）；跑一整夜智能体任务，也就一杯咖啡

钱花在哪



这份 PPT , 全程约 2 万 token

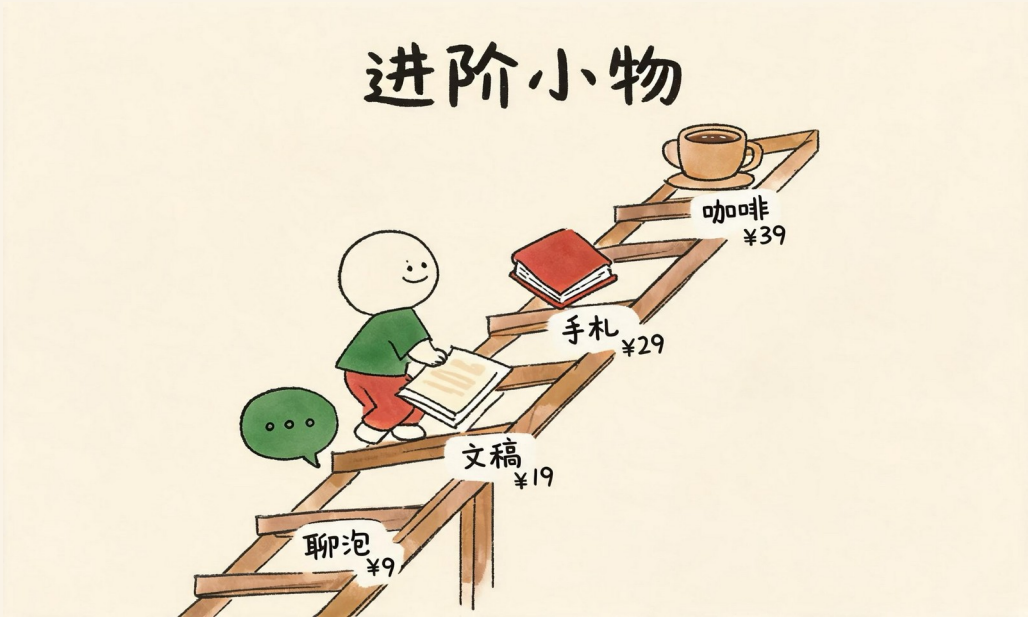
生态



- 大纲 1 万字 + 63 页文字 + 给 AI 的指令 , 总共约 2 万 token
- 用 Claude Sonnet (\$3 / \$15) $\approx$ 1 毛人民币
- 用 DeepSeek 更便宜——一次讲座 , 几分钱

现场可指着这份 PPT : 就是你现在看的

任务	Token 量	成本
问一句、改一段	几千	几乎不要钱
审一篇论文 (约 1 万字)	~1 万	几分钱
读一本书 (约 20 万字)	~20 万	几块钱
跑一整夜智能体任务	几百万	一杯咖啡





01 通用自主智能体

给目标自动拆解、交交付物；过程偏黑盒

02 低代码平台

拖拽搭建、一键发布到飞书或微信

03 代码框架

LangGraph、AutoGen、CrewAI，深度定制

看三件事：数据能不能出去、要不要私有化、有没有人手

场景怎么选：科研 · 办公 · 编码

生态

科研

文献检索与核实、数据整理、写作辅助——重可靠性

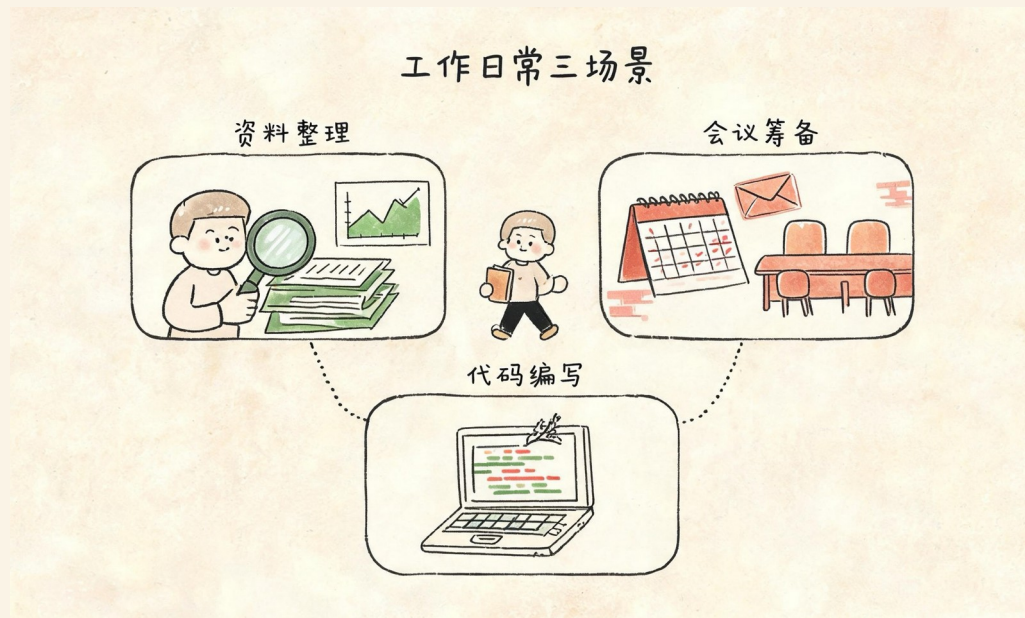
办公

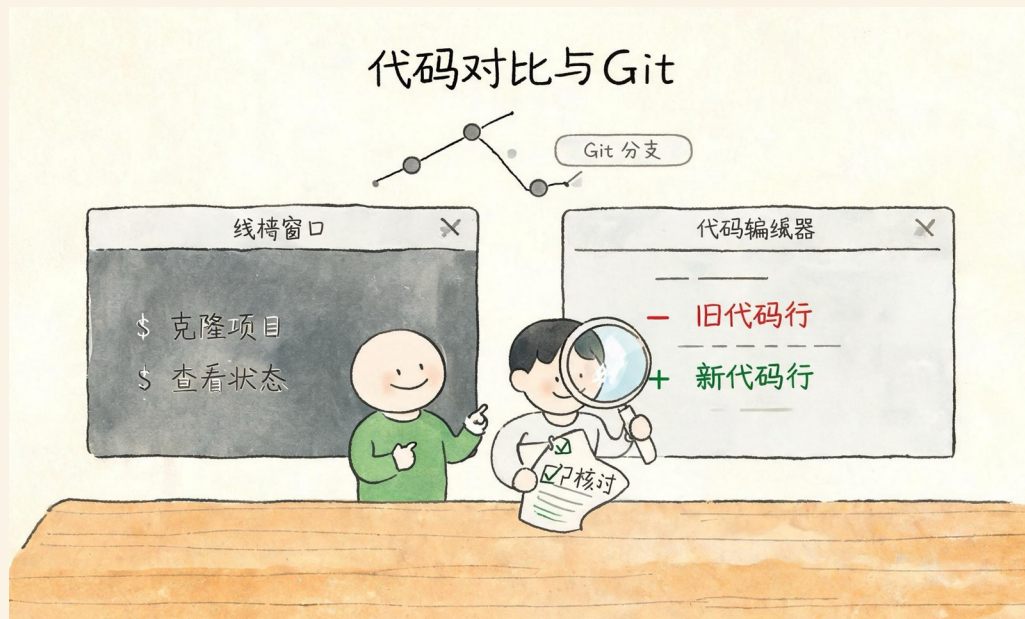
会议纪要、文档流转、跨系统自动化——重集成度

编码

直接改写代码库的编码智能体——重工具链与控制权

共同点：都要能落到你已有的文件与系统里，否则只是玩具





- **Claude Code**：跑在终端里，直接读写你的代码库，适合命令行工作流
- **Cursor**：AI 原生的编辑器，跨文件重构与改写，人和它在同一个界面里干活
- **GitHub Copilot**：最早的补全式助手，生态最广，多数编辑器都能装
- 共同点：它能改真实文件——所以版本控制与逐次过目是前提

别在没备份的仓库上放开手；让它改完，你仍要看 diff

PART 03

以小荷为依托：安装、配置与使用



从装到用，走一遍完整流程



- 自托管的个人 AI 助理平台
- 多模型接入：DeepSeek、Kimi、Claude 等自由切换
- 技能可扩展：学术、文档、绘图、配音、视频
- 一套会话，终端、浏览器、桌面、手机**同一份记忆**

把“云端助理”搬回你自己的机器上，还能接你想用的模型

架构总览：五个部分各管什么

实操

agent

服务端与内嵌网页，系统的中枢

client

桌面与手机端，你看到的界面

llm

模型网关，统一各家模型调用

mcp

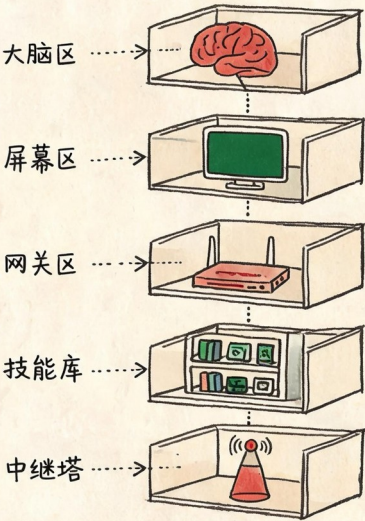
技能 Hub，技能发现与分发

derp

组网中继，跨网络也能连

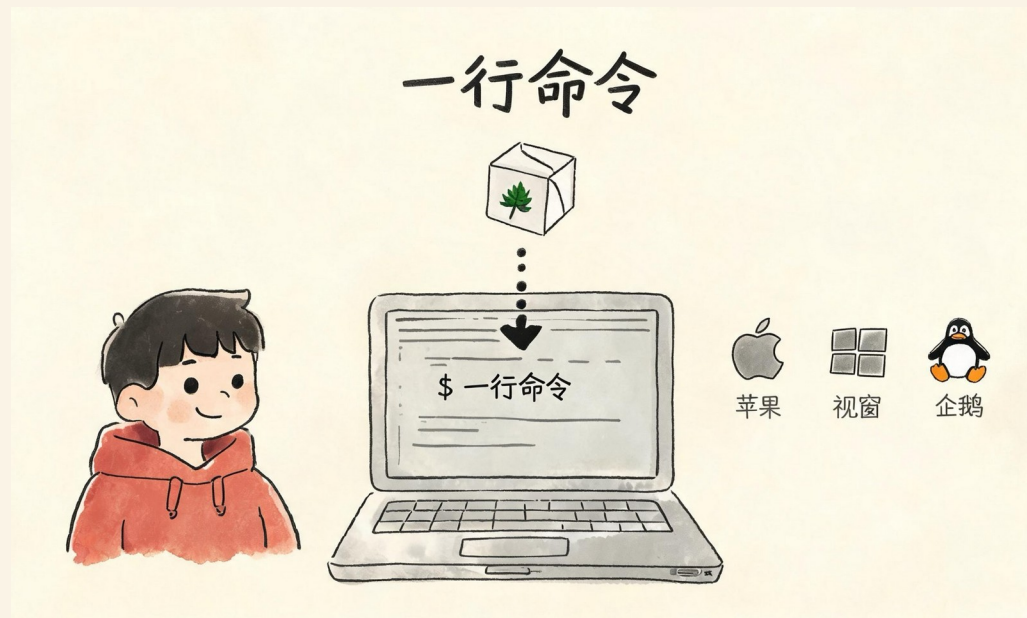
单机自用只需要 agent 与 client，其余按需接入

五个部分



安装：一行命令，三个平台

实操



macOS / Linux

终端里一行命令安装，自动配置环境与依赖

Windows

提供安装包，双击即可，不需要先装 Python

装完自带升级与回滚，保留最近两版，随时可退回

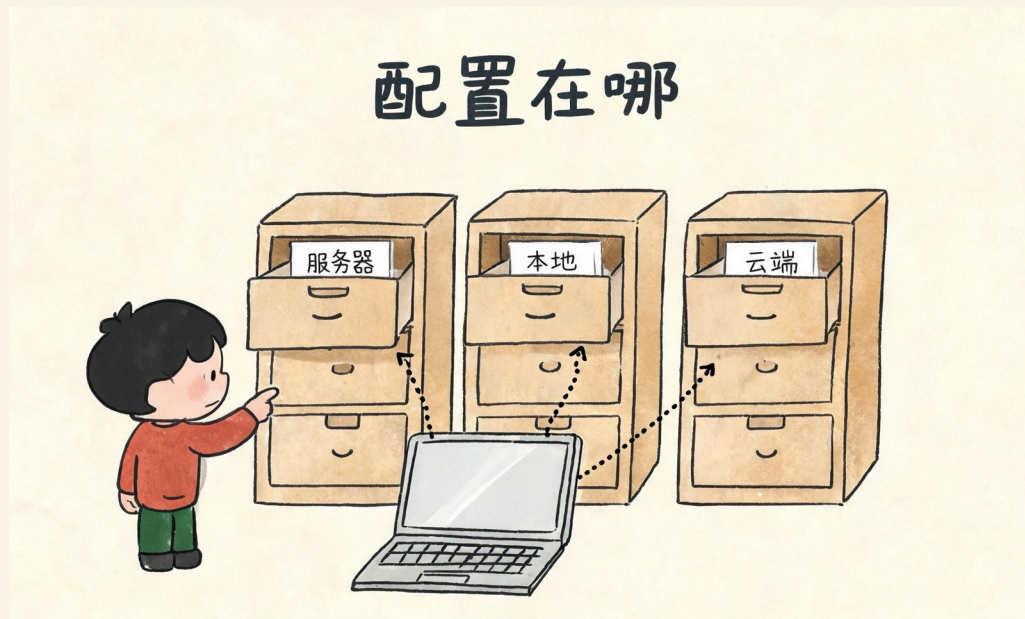
首次配置：接上你的模型

实操

- 01 到模型厂商官网申请一个 API Key
- 02 在小荷里选模型提供商，粘贴 Key，指定默认模型
- 03 语音、飞书等能力按需再开，不影响主流程

也可以完全用本地模型——一个 Key 都不填，照样能跑





技能与工具类密钥
`~/.mytoolkit/config.json`

模型服务商密钥
`~/.xiaohe/agent/secrets.json`

运行设置
`~/.xiaohe/agent/settings.json`

密钥缺失时它会主动问你要，并写进对应位置

界面导览：一个会话，四种入口

实操

终端

命令行直接对话，适合批处理与脚本化

浏览器

打开网页就能用，不用装东西

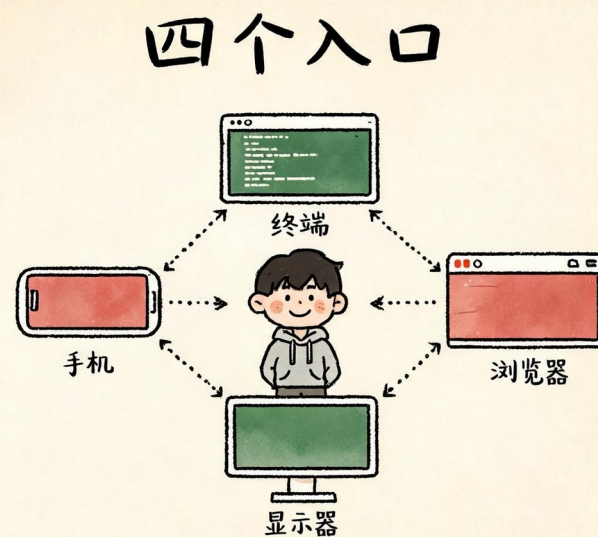
桌面客户端

常驻菜单栏，随手唤起

手机

同一份会话继续，通勤路上接着看

四个入口共用同一套技能与同一份记忆





智能体讲座 /

你打开的就是这一层

AI 智能体讲座 .pptx

成品，双击就能放

README.md

写了结构、怎么改

00- 参考资料 /

复用的三份旧 deck

deck/

它的工作台

outline.yaml

大纲：文字的真源

slides/

每页一个 HTML

images/

配图

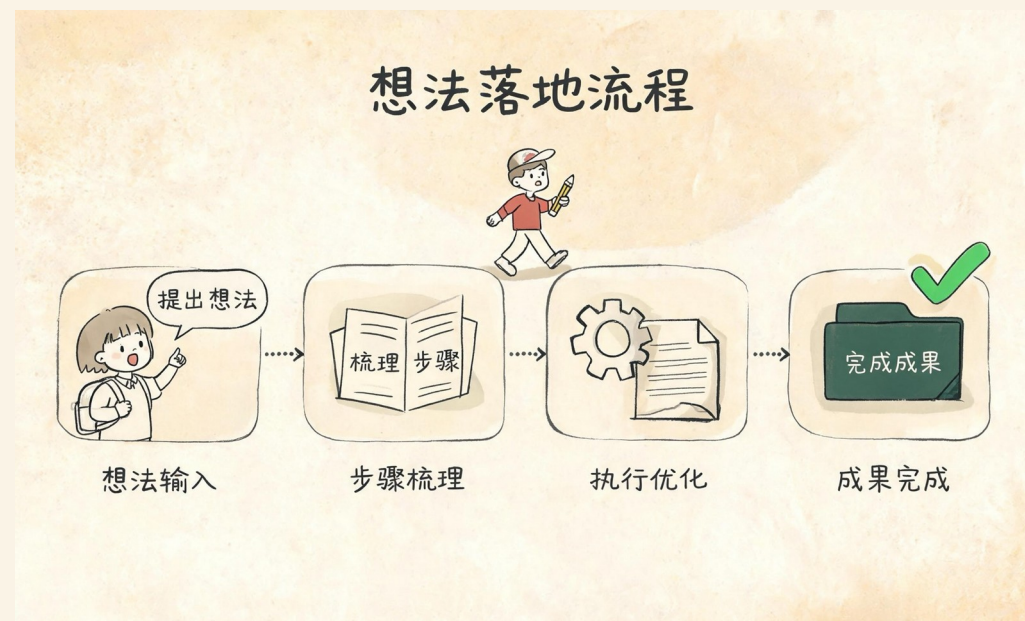
看得见、点得开、改得动——**不是黑盒**

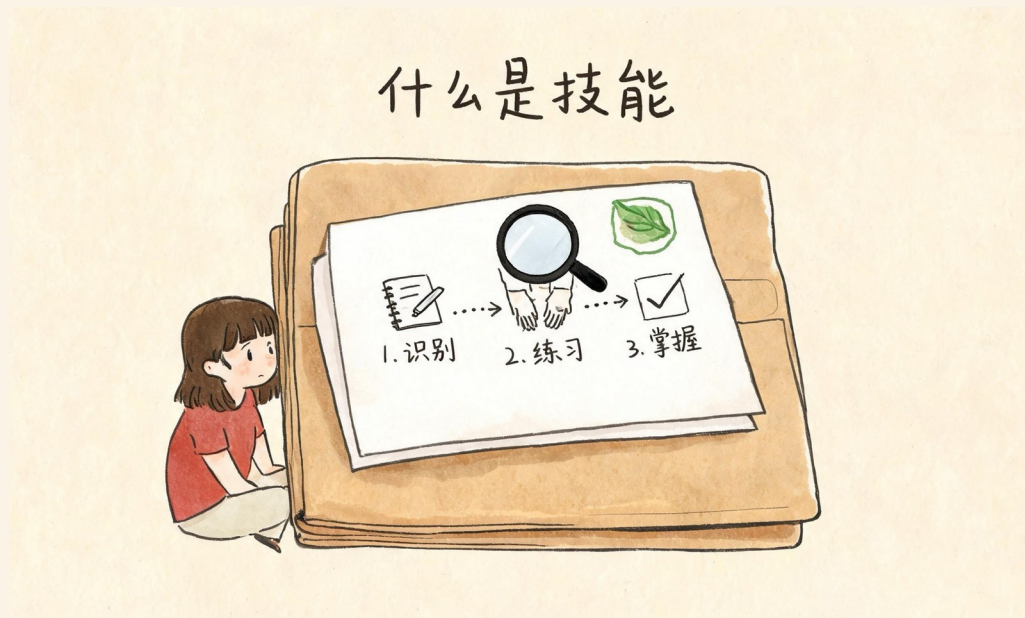
一次会话是怎么跑的

实操

你说：“把第 39 页那张图换掉”

- 01 读约定** 先读 AGENTS.md，知道你的目录怎么放、命名什么规矩
- 02 找文件** 只在工作区里找；工作区之外它看不见
- 03 按技能做** 匹配到对应技能，照技能里写的步骤走
- 04 边做边报** 改哪个文件都告诉你；拿不准就停下来问你
- 05 落回文件** 产出进工作区，结论写回文件，下次不用重说





- 技能 = 一个文件夹 + 一份 **SKILL.md** 说明书
- 说明书告诉助理：什么时候用、按什么步骤做、产出什么
- 平时只加载一句话摘要，用到时才读全文
- 装技能 = 放进目录；不用了移走，不改主程序

本质是把“老师傅的做法”写成文档，让助理照着做

三样东西，各管一段

实操

AGENTS.md

长期约定——每次都遵守的口径与规矩

技能 Skill

一类任务的做法——遇到才加载

记忆

攒下来的事实——上次的结论、你的偏好

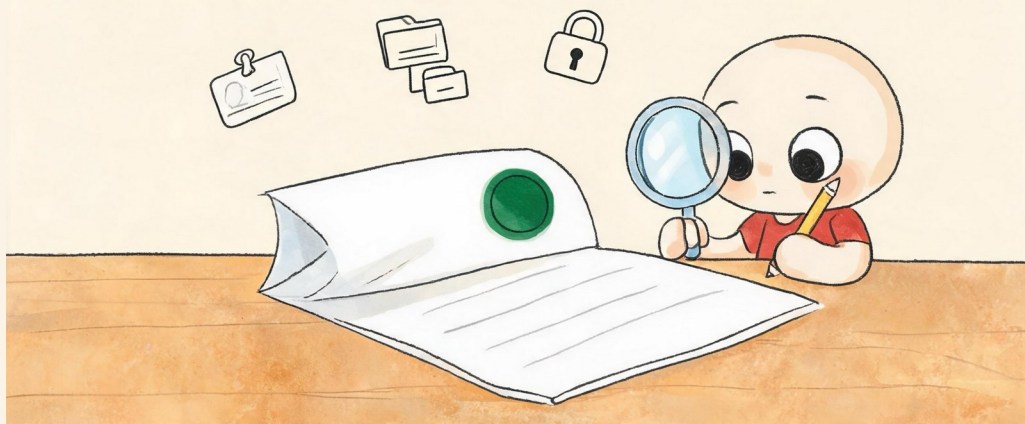
配置

密钥与开关——接哪个模型、开哪些能力

四样东西



先读这份

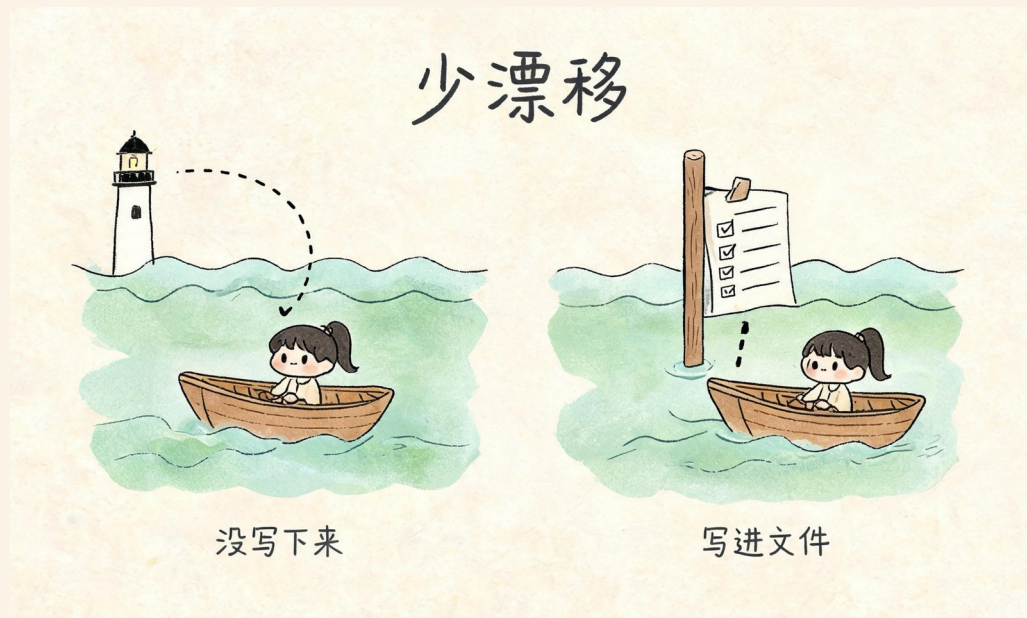


- 每次会话开始，它**先读的就是这份文件**
- 里面写：你是谁、怎么称呼、用什么语气、目录怎么放、文件怎么命名
- 也写禁区：哪些不能做、哪些要先问过你
- 它管的是“长期一直遵守”，不是这一件事怎么做

技能解决“这件事怎么做”，AGENTS.md 解决“你一贯要什么”

- 三种典型表现：称呼口径变了、文件乱放、同一件事反复解释
- 根因：**没写下来的要求，它每个会话都得重新猜一遍**
- 维护只有两个动作：写进 AGENTS.md，或固化成技能

一条准则：凡是说过第二遍的要求，就该写进约定文件





学术研究

文献检索与核实、审稿意见、基金申报写作

文档产出

Markdown 一条命令转 Word / PDF

绘图渲染

数据图表、结构示意图、三维出图

配音视频

逐页配音，直接合成讲解视频

办公事务

浏览器自动化、日程与邮件、资料整理

案例一：文献检索与核实

实操

01 问题

模型写出的参考文献，有时根本不存在

02 做法

按题名逐条回查数据库，而不是直接采信

03 产出

每条标注“已核实”或“未找到”

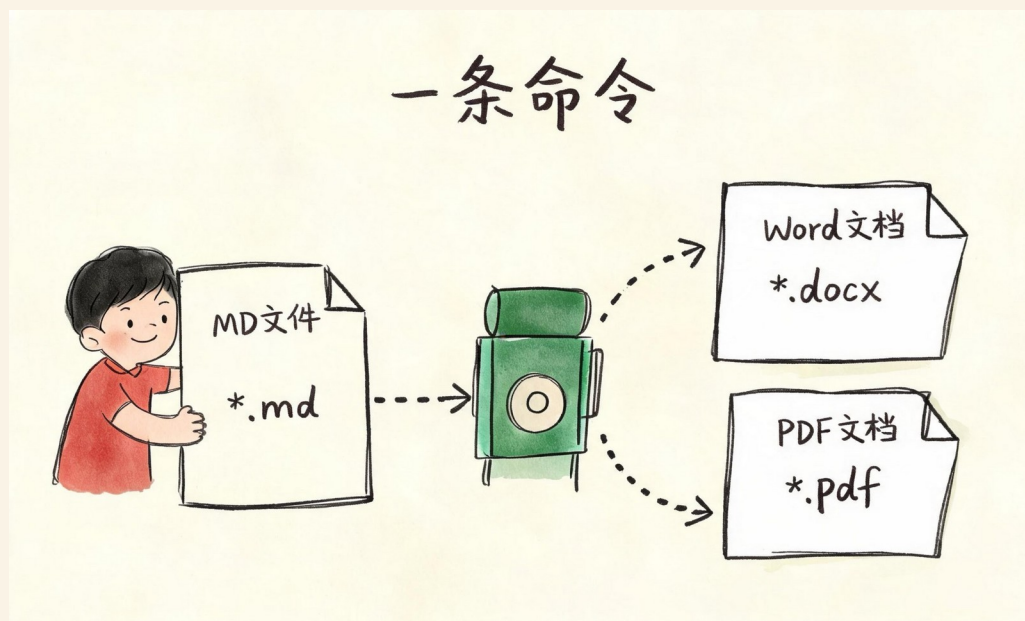
查不到不等于造假，但必须逐条过一遍

逐条核实



案例二：Markdown 一条命令出文档

实操



- 写 Markdown 源文件，一条命令转出 Word 与 PDF
- 公式转成 Word **原生公式**，可直接复制编辑
- 中文排版（缩进、行距、图注、三线表）自动处理

```
mytoolkit convert 报告.md --to docx  
mytoolkit convert 报告.md --to pdf
```

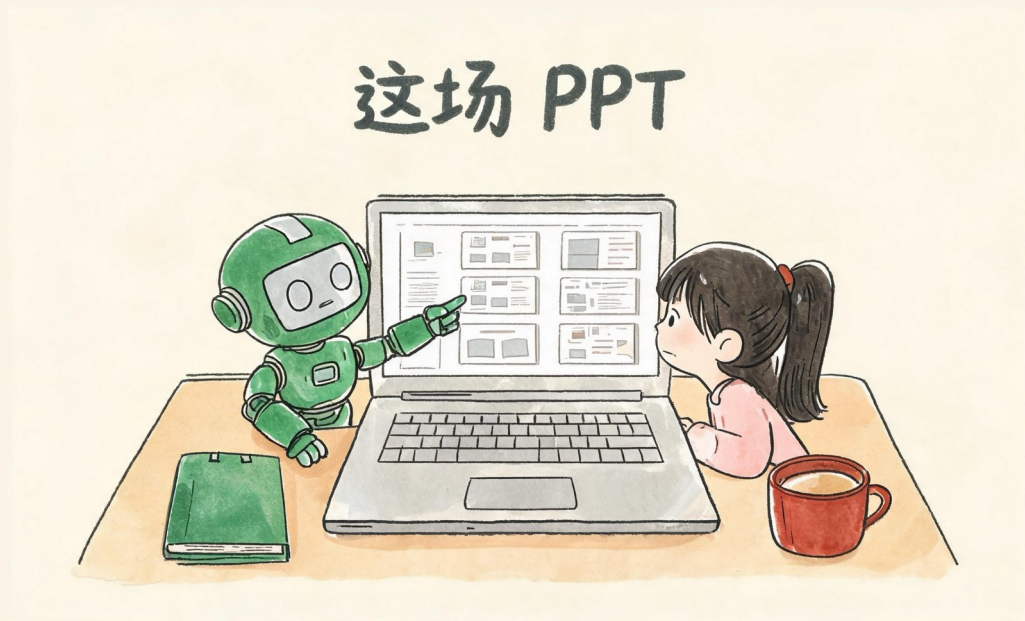
案例三：这场讲座的 PPT 是怎么来的

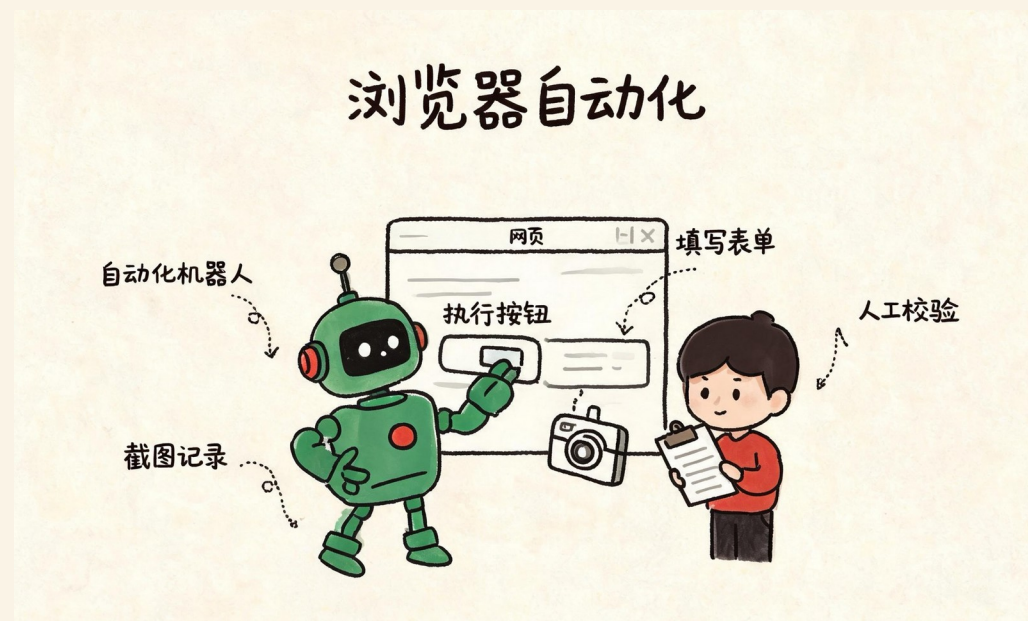
实操

定需求 → 定风格 → 出大纲 → 逐页生成 → 导出

- 每页是一个**独立 HTML**，转成矢量图后打包为 PPTX
- 文字在 PowerPoint 里**可编辑**，不是图片
- 素材、大纲、成品都留在同一个项目文件夹里

你现在看的这份，就是用这套流程做出来的



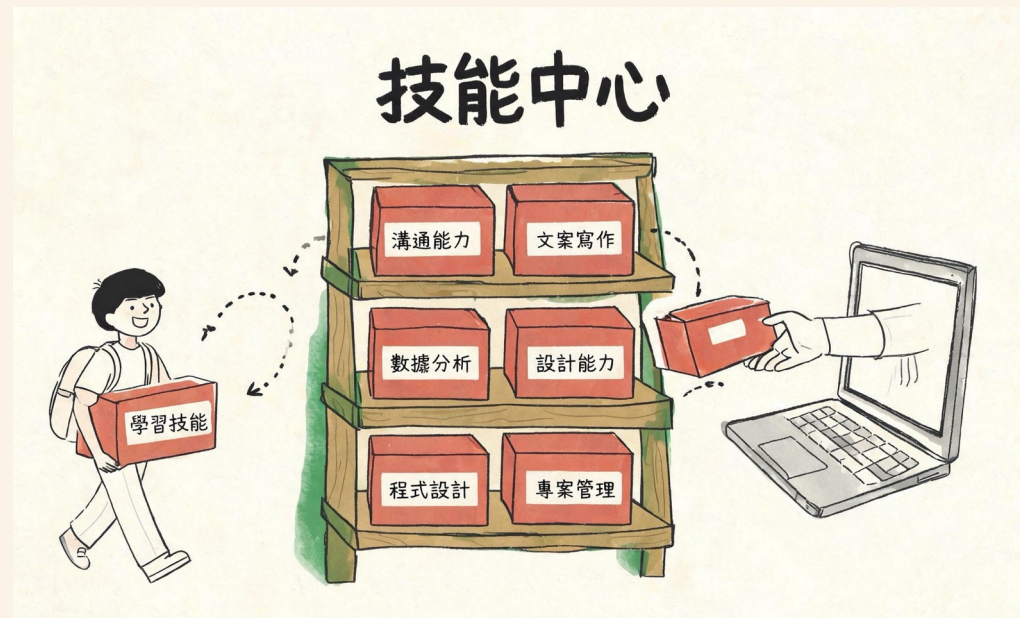


- 驱动真实浏览器登录、填表、上传、截图
- 适合**没有开放接口**的校内系统与后台
- 每一步截图确认，避免批量误操作

敏感操作保留人工确认，不自动提交——这条边界必须守住

- Hub 是技能的集中分发处，按需安装到本地
- 渐进披露：摘要先加载，用到才读全文
- 版本可更新，旧版留在本地可回退
- 团队可以维护自己的技能集，**一次分发全体可用**

技能也可以被版本管理和审阅，就像代码



扩展自己的技能：从重复劳动开始

实操

写自己的

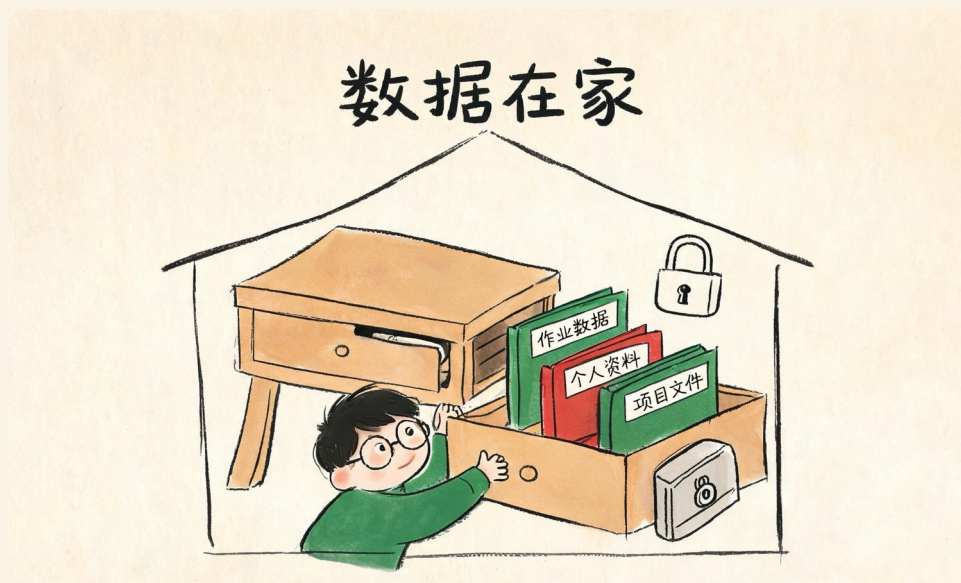


01 挑一件每周都要做、步骤固定的事

02 把步骤写下来，就是 SKILL.md 的初稿

03 放进技能目录，让它跑一遍，再逐次修正

好技能的标准：别人拿去也能跑通



- 会话与文件落在**自己的机器与服务器**上
- 模型调用按需直连，不经第三方托管中转
- 工作区就是你的文件夹：看得见、管得住、能备份

对涉密材料与未发表成果，这是“能不能用”的前提

数据 成本 排错



数据在哪

工作区文件夹、本机配置目录、自己的服务器

成本怎么控

按任务选模型、开缓存、长任务拆小步

出错怎么办

先看日志，会话与运行日志都在本机

怎么备份

工作区就是文件夹，直接复制或同步

一条底线：涉密材料不进公网模型——进了，就要当作已经公开

原理

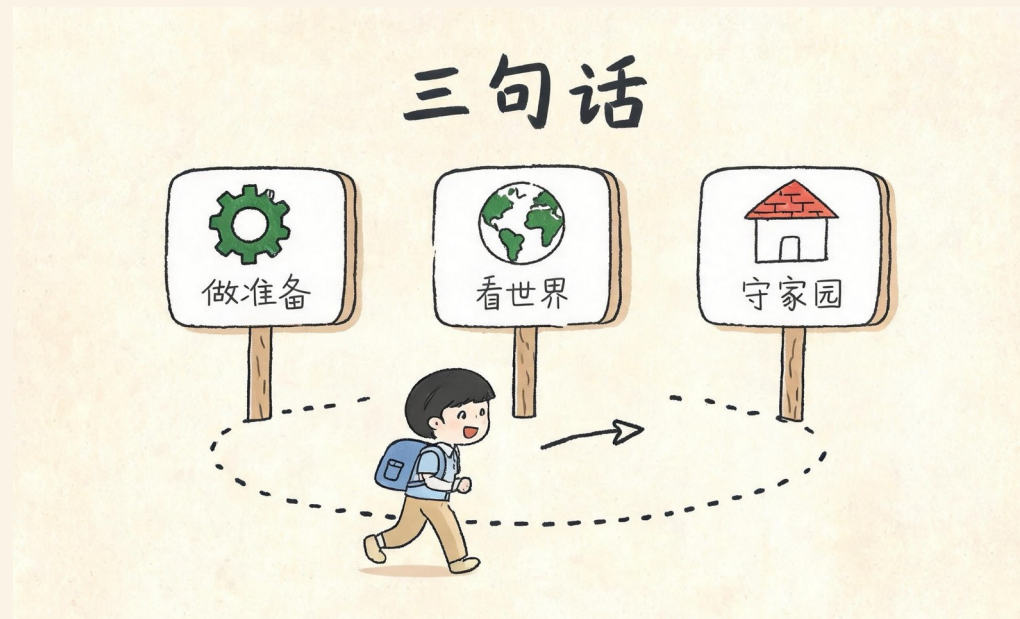
大模型是“预测下一个词”的引擎，智能体在它外面装上了手脚与循环

生态

海外三强拼能力，国产拼性价比，开源权重决定数据能否留在内网

实操

自托管这条路是通的——数据在你手里，能力按需接上



谢谢！敬请批评指正。

AI 智能体讲座

AI 智能体 · 从大模型到你的个人助理