



每个人都可以听懂的大模型科普讲座

智能体OpenClaw (小龙虾) 应用实践

林子雨 副教授

厦门大学

2026年3月16日



大模型系列报告（科普报告，非技术报告）

全网浏览量远超1000万

- 报告1：大模型概念、技术与应用实践（面向社会大众）
- 报告2：DeepSeek大模型赋能高校教学和科研（面向高校）
- 报告3：DeepSeek大模型及其企业应用实践（面向企业）
- 报告4：DeepSeek大模型赋能政府数字化转型（面向政府部门）



扫码下载报告

报告下载地址：<https://dbllab.xmu.edu.cn/post/deepseek/>



大模型系列讲座

2025年3月至今已完成130场讲座

厦大台湾研究院、厦大管理学院、厦大信息学院、厦大管理学院MBA、中山大学商学院MBA
厦大法学院、厦大全体辅导员、自然资源部海洋三所、公安部一所、厦门国家火炬高新区、厦门律师行业
厦门思明区公务员、全国高职院校教师培训班、厦大教师发展中心、300多家制造企业AI培训班
龙岩为龙净环保、厦门IT联盟、厦门银城智谷园区企业、瑞安市科级干部、厦门广电网络、浙江工商大学
福建省农业农村厅、福建省市场监督管理局、晋江市房屋建设公司、华为山西分公司、全国高校直播、天津大学



厦门大学大数据教学团队



林子雨



夏小云



赖永炫



陶继平



谢怡



郑海山



蔡炳跃



苏淑文



郑宇辉



张琦

国内高校大数据教学的重要贡献者

团队负责人：**林子雨** 副教授

年轻力量：核心成员全部46周岁以下

结构合理：教学型、科研型、实验工程师

专注专业：从2013年至今，11年专注于大数据教学

团队特点：眼光前瞻、紧跟技术、创新实干、执行力强

影响力高：多项指标在国内高校大数据教学领域领先

- 教材数量
- 教材占有率
- MOOC课程学习人数
- 师资培养
- 教学研讨会
- 教学网站访问量
- 在线讲座观看人数
-

团队联系方式：ziyulin@xmu.edu.cn

厦门大学大数据教学团队

20余本大数据与人工智能教材服务全国1000多所高校



教材详情: <https://dbl原因.xmu.edu.cn/post/bigdatabook/>

目录

1. 人工智能发展简史
2. 人工智能思维
3. 详解OpenClaw
4. OpenClaw的云端部署和应用
5. OpenClaw辅助科研



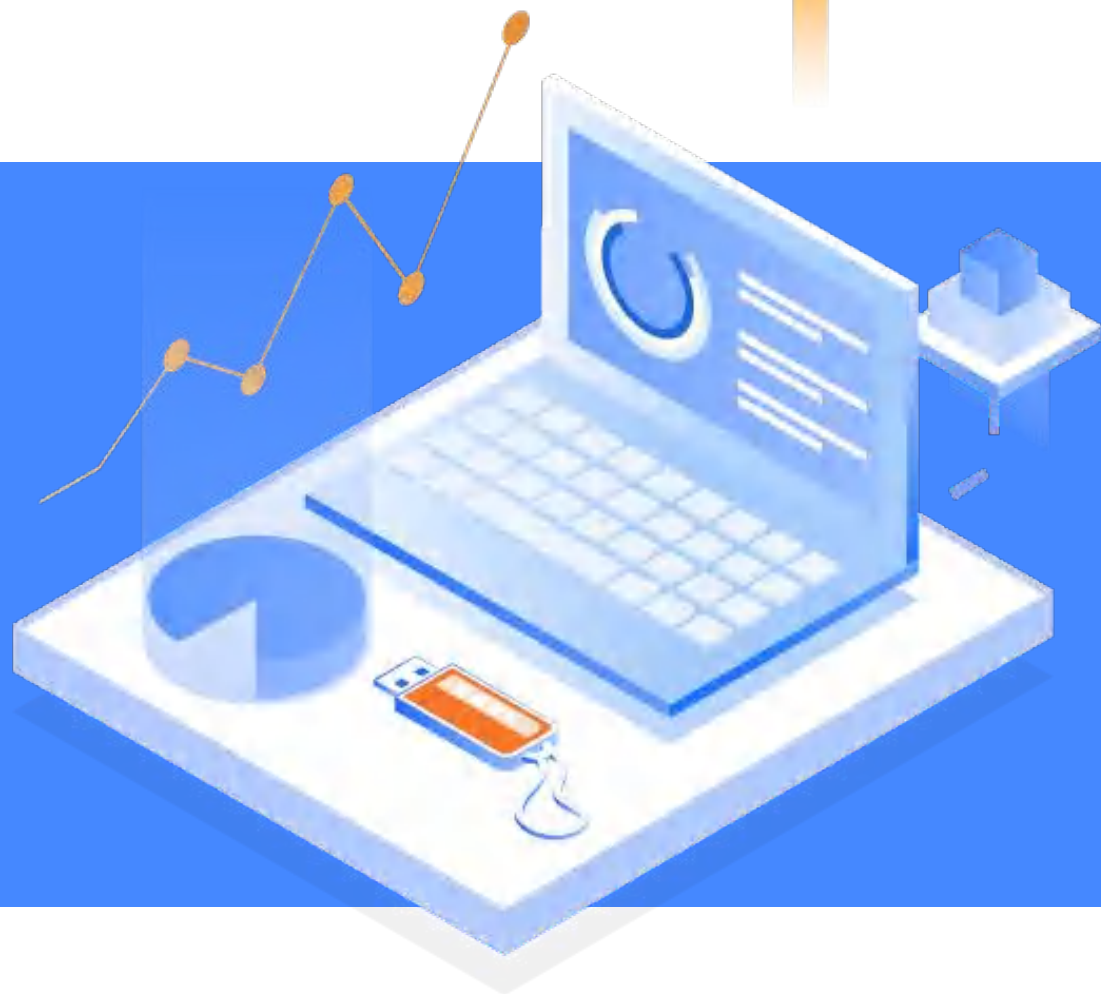
1. 人工智能发展简史

1.1 图灵测试

1.2 人工智能的诞生

1.3 人工智能的发展阶段

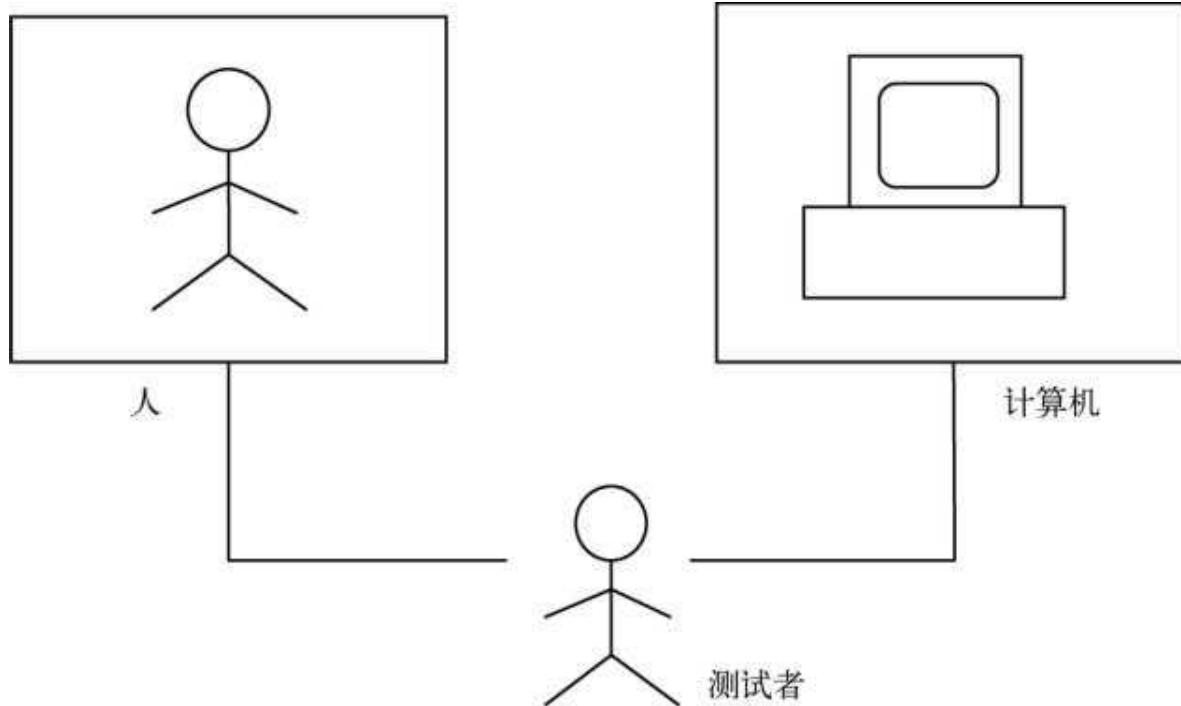
1.4 未来人工智能发展的五个阶段





1.1 图灵测试

1950年，“计算机之父”和“人工智能之父”艾伦·图灵（Alan M. Turing）发表了论文《计算机与智能》，这篇论文被誉为人工智能科学的开山之作。在论文的开篇，图灵提出了一个引人深思的问题：“机器能思考吗？”。这个问题激发了人们无尽的想象，同时也奠定了人工智能的基本概念和雏形



在这篇论文中，图灵提出了鉴别机器是否具有智能的方法，这就是人工智能领域著名的“图灵测试”。如图所示，其基本思想是测试者在与被测试者（一个人和一台机器）隔离的情况下，通过一些装置（如键盘）向被测试者随意提问。进行多次测试后，如果被测试者机器让平均每个测试者做出超过30%的误判，那么这台机器就通过了测试，并被认为具有人类智能



1.2 人工智能的诞生

人工智能的诞生可以追溯到20世纪50年代。当时，计算机科学刚刚起步，人们开始尝试通过计算机程序来模拟人类的思维和行为。在这个背景下，一些杰出的科学家和工程师们开始研究如何使计算机具备更高级的功能

1956年8月，在美国达特茅斯学院举办的人工智能夏季研讨会，是人工智能领域具有里程碑意义的一次重要会议。这次会议汇聚了众多杰出的科学家和工程师，他们共同探讨和研究人工智能的发展和应用前景

这次会议的主题围绕着**人工智能的定义、研究方法和应用场景**展开。与会者们深入探讨了人工智能的基本概念、算法和技术，以及其在各个领域的应用潜力。他们共同认识到，人工智能的研究和发展将为人类带来巨大的变革和进步

1956 Dartmouth Conference:
The Founding Fathers of AI



John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester



Trenchard More





1.2 人工智能的诞生

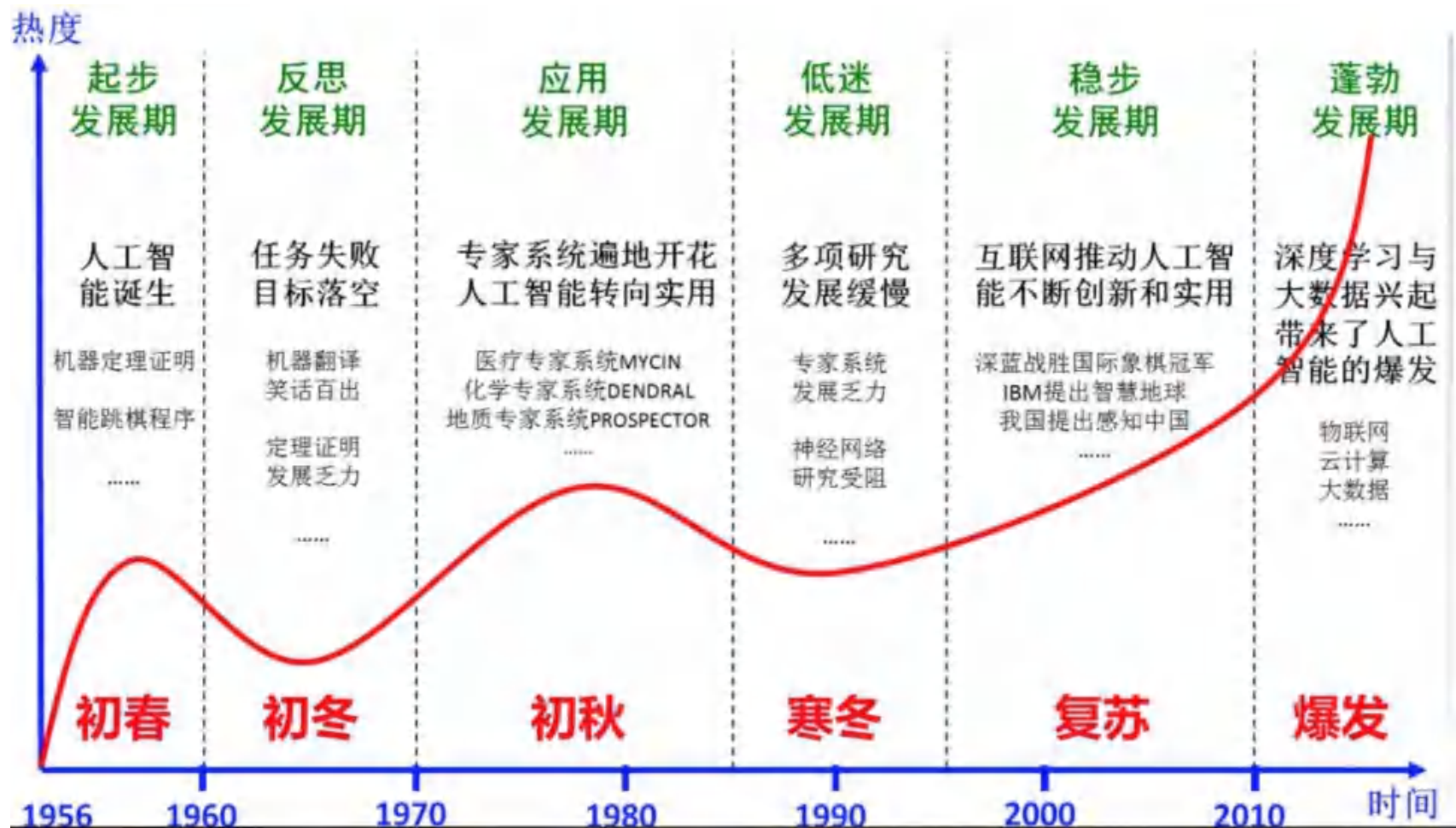
在这次会议上，“人工智能”这个词汇被约翰·麦卡锡（John McCarthy）首次提出。与会者们不仅对人工智能的研究和应用前景进行了深入探讨，还提出了许多重要的观点和思路，为人工智能的发展奠定了基础。**这次会议的召开标志着人工智能作为一个独立学科的正式诞生，因此，达特茅斯会议被称为“人工智能的开端”，1956年也被称为“人工智能元年”。**这次会议不仅为人工智能的研究和发展奠定了基础，还为人类带来了巨大的变革和进步





1.3 人工智能的发展阶段

从1956年人工智能元年至今，人工智能的发展历程经历了漫长的岁月，大致可以划分为以下6个阶段





1.4 未来人工智能发展5个阶段



OpenAI 的 5 级 AGI 量表

1-5 级说明

- L5: 可以完成整个组织工作的人工智能。
- L4: 可以创造新事物的人工智能。
- L3: 可以代表用户采取行动的人工智能代理。
- L2: 像博士一样解决问题，无需工具。
- L1: 当前AI，类似ChatGPT，可以与人类对话的人工智能。





1.5 AI能力的四层金字塔

感知层

第一层：感知 Perception
看懂文字、图片、语音、视频。这是基础能力层。2024年基本解决。



认知层

第二层：认知 Cognition
理解含义、逻辑推理、知识关联。大模型的核心能力。2023-2025年快速进化。



Logic and Reasoning

决策层

第三层：决策 Decision
在多个选项中选择最优方案、制定执行计划。o1/DeepSeek-R1在这一层取得突破。



行动层

第四层：行动 Action ← OpenClaw
在这里将决策转化为物理或数字世界的实际操作—调用工具、操作系统、运行代码、发送通讯。这是2026年正在突破的层。OpenClaw就是打通这一层的基础设施。





1.6 大模型的能力边界

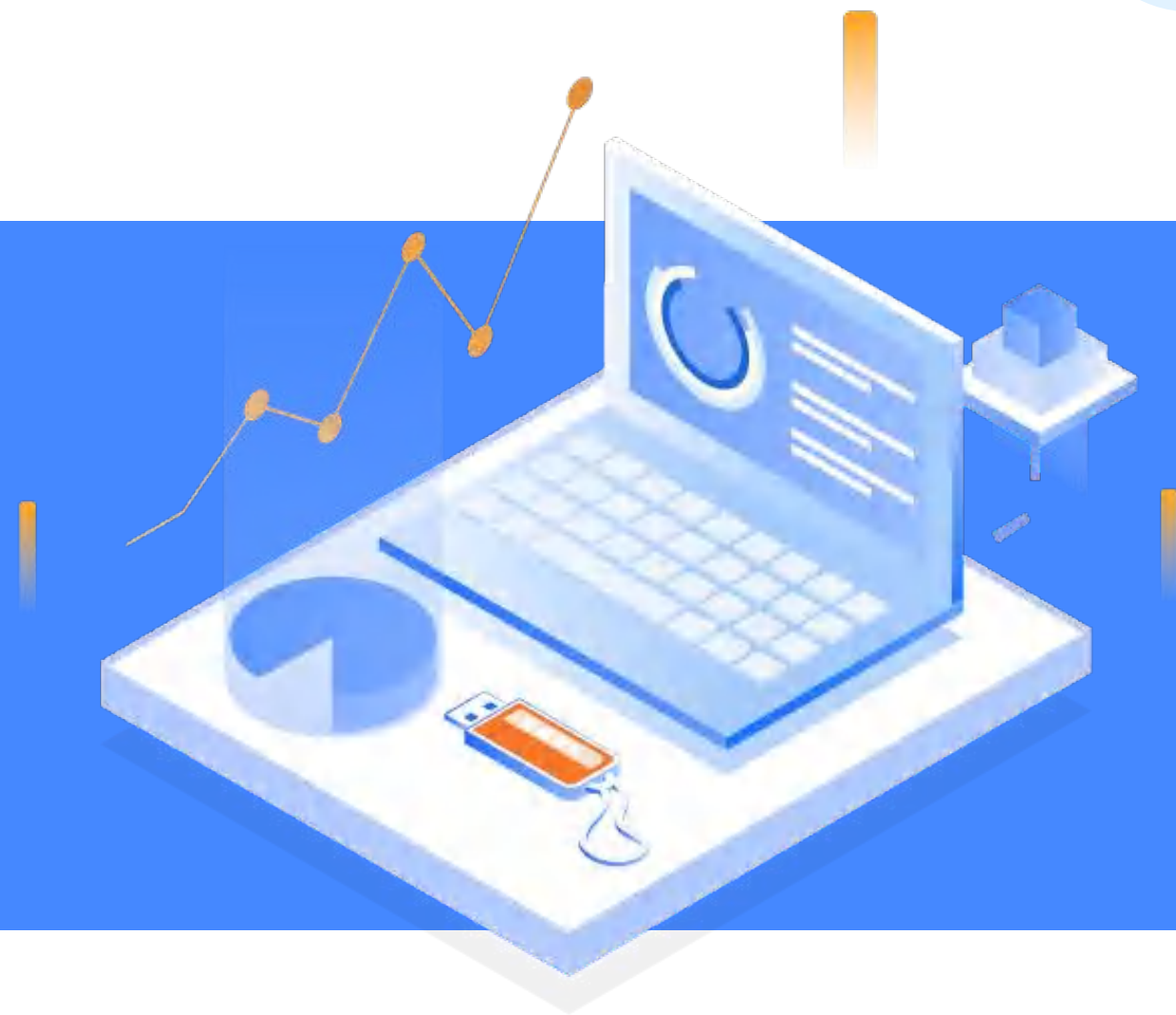
能力维度	当前水平 (2026年3月)	你的应对策略
文本理解与生成	极强。超越 95%人类写作者	放心用，但核验事实
逻辑推理	强。复杂数学和编程已接近专家	用于辅助，但审查推理链
知识广度	极广。但有知识截止日期	实时信息必须联网搜索验证
事实准确性	不可靠。会自信地胡说八道	所有事实必须交叉验证
创造性	中等。善于组合，弱于突破	用于素材生成，原创靠你
情感理解	表面。能识别情绪，不能真正共情	涉及人际的事必须人工判断
价值判断	极弱。没有真正的价值观	永远不要让 AI 替你做价值判断
执行能力 (代理)	初级。能做简单任务链	从低风险任务开始，逐步信任



1.7 未来3-5年的发展趋势

时间窗口	大概率发生	对你的影响
2026-2027	<ul style="list-style-type: none">✓ AI 代理从极客玩具进入主流商业场景✓ 多代理协作成为现实	先行者建立认知优势和数字资产
2027-2028	<ul style="list-style-type: none">✓ 代理间标准化通讯协议成熟✓ 个人 AI 助手成为标配	不会用 AI 代理的人开始感到明显劣势
2028-2030	<ul style="list-style-type: none">✓ AI 代理深度嵌入企业流程✓ 「AI 原生」公司涌现	组织架构和职业分工发生根本性重塑

2. 人工智能思维





2. 人工智能思维



了解

每个人都应了解人工智能
的基础运行模式



区分

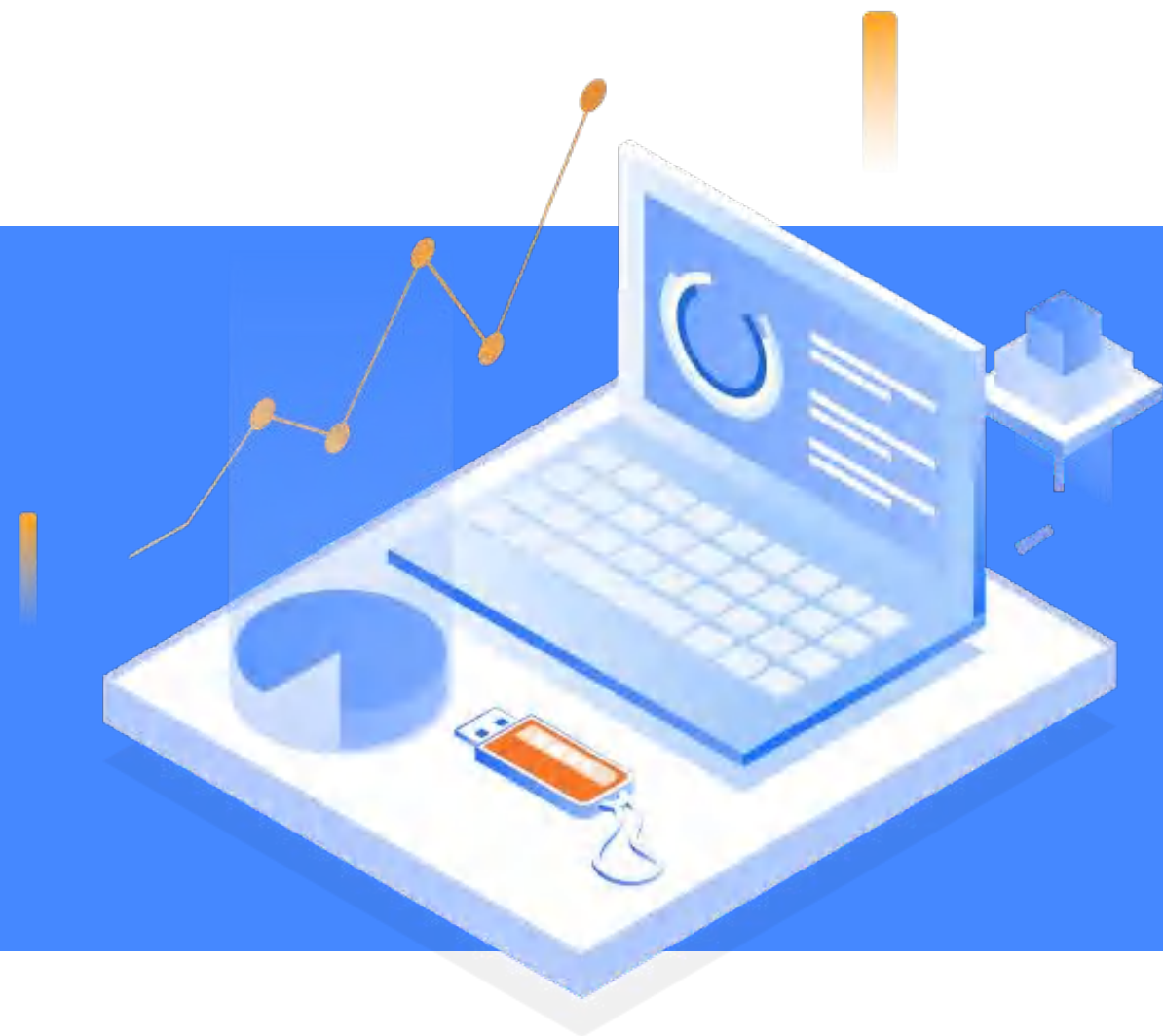
具备区分人的能力
和机器的能力



协作

拥有和人工智能协作的能力，
懂得如何运用人工智能

3. 详解OpenClaw





3.1 什么是OpenClaw(小龙虾)

项目名称

OpenClaw

开发者

奥地利独立开发者Peter Steinberger (彼得·斯坦伯格), PSPDFKit创始人

更名历史

2025年11月创建时名为Clawdbot, 2026年1月27日因商标纠纷更名为Moltbot, 2026年1月30日正式定名为OpenClaw

相关事件

2026年2月, Peter Steinberger加入OpenAI团队





3.1 什么是OpenClaw(小龙虾)

OpenClaw

OpenClaw是一个开源AI智能体执行网关，它连接大模型的智能决策能力与本地设备的执行功能，通过自然语言指令实现AI自动完成复杂任务，显著提升工作效率和自动化水平



功能对比

- 传统 AI:** 告诉你 “怎么整理文件”
- 龙虾 AI:** 直接帮你把几千个文件分好类、重命名、归档
- 传统 AI:** 给你写周报大纲
- 龙虾 AI:** 打开文档、填数据、生成完整周报、发到群里

OpenClaw开源AI智能体



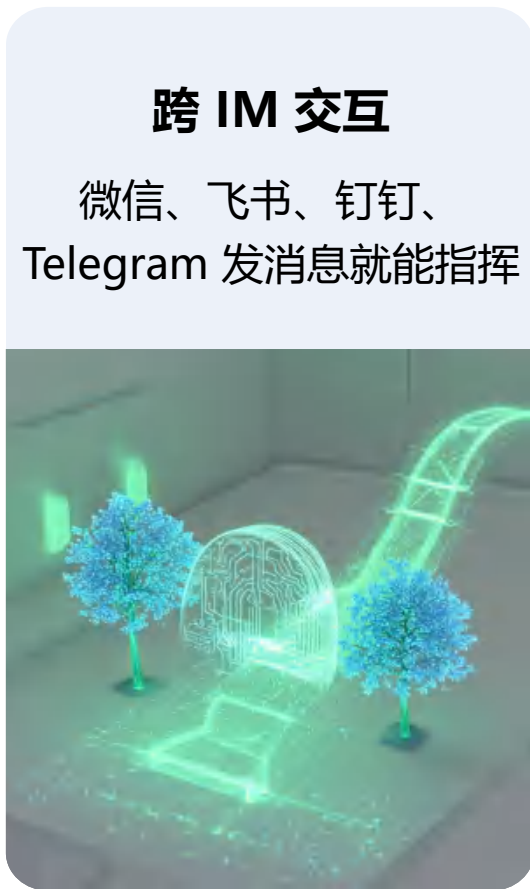
3.1 什么是OpenClaw(小龙虾)

OpenClaw 核心能力



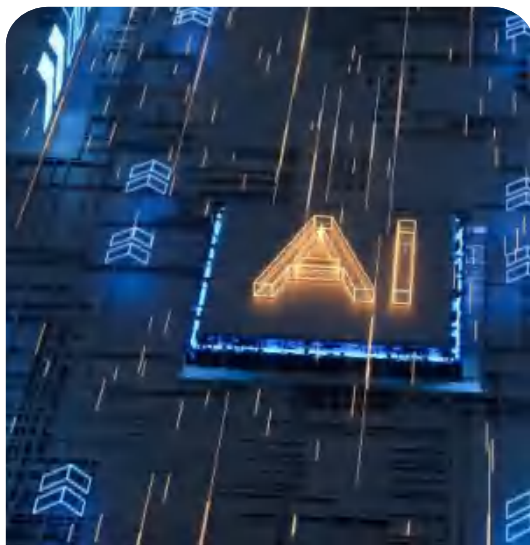
本地执行

直接操作电脑文件、
浏览器、软件



跨 IM 交互

微信、飞书、钉钉、
Telegram 发消息就能指挥



持久记忆

记住你的习惯、历史任
务，越用越懂你



技能扩展

像装 App 一样安装行
业技能，变成专业助理



3.1 什么是OpenClaw(小龙虾)

核心维度	能力描述	价值体现
端到端任务执行	自主规划并完成复杂流程（如：自动处理邮件→生成报告→部署代码）	拒绝空谈： 不仅理解需求，更直接交付结果
系统级集成能力	无缝对接本地应用、云服务API、数据库及物联网设备	打破孤岛： 连接碎片化系统，实现数据自由流动
多智能体协同	支持Manager/Worker架构，构建虚拟专家团队并行处理任务	效率倍增： 解决记忆冲突，实现高并发下的稳定执行
持久化记忆系统	结合RAG技术，拥有可搜索的长期记忆与个性化偏好	越用越懂你： 具备上下文连续性，实现真正的“第二大脑”

OpenClaw核心能力引擎



3.1 什么是OpenClaw(小龙虾)

典型应用场景

场景分类	核心功能	真实案例/实例
智能助理	自动读取邮件、提取会议信息、同步日历、发送定制消息	猎豹移动CEO傅盛 : 夜间自动向600+联系人发送定制拜年信并安排行程
第二大脑	抓取 Telegram/X/YouTube/PDF 内容, 向量化存储, 支持语义检索	Matthew Berman : 构建可自然语言查询的个人知识库, 自动归档所有阅读内容
健康管理	拍照识别食物、记录饮食、关联症状分析、用药提醒	食物日记系统 : 通过交叉分析发现潜在过敏源, 形成健康闭环
商业顾问团	8个AI专家角色 (财务/营销/运营等) 并行分析14+数据源, 输出决策建议	一人公司智囊团 : 每天凌晨自动生成按优先级排序的业务建议清单
安全委员会	每晚审查代码库、Git历史、运行日志, 自动修复漏洞	AI自查系统 : 4个维度安全专家并行工作, 关键问题即时告警
内容创作流	从Slack捕捉创意→全网调研→查重→生成大纲→创建Asana任务	视频选题流水线 : 全自动完成从灵感到项目立项的全过程



3.1 什么是OpenClaw(小龙虾)

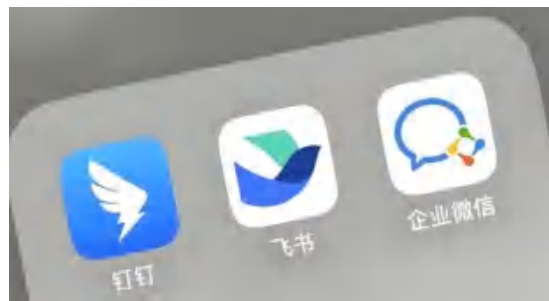
核心业务赋能

智能办公中枢



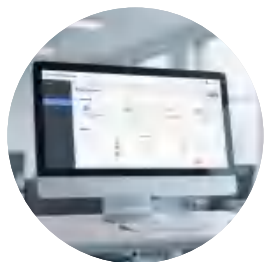
统一入口：集成飞书/钉钉/企业微信

@机器人即可触发跨系统操作



多Agent协同

采用 Manager-Worker 架构（如阿里云 HiClaw），任务拆解与内存隔离，解决复杂任务中的“记忆爆炸”问题



成效

显著降低人工干预，实现高管日程管理与项目任务自动编排





3.1 什么是OpenClaw(小龙虾)

核心业务赋能

客户服务自动化



全渠道响应：接入微信/网页/电话/邮件，
7×24小时自动应答、生成工单

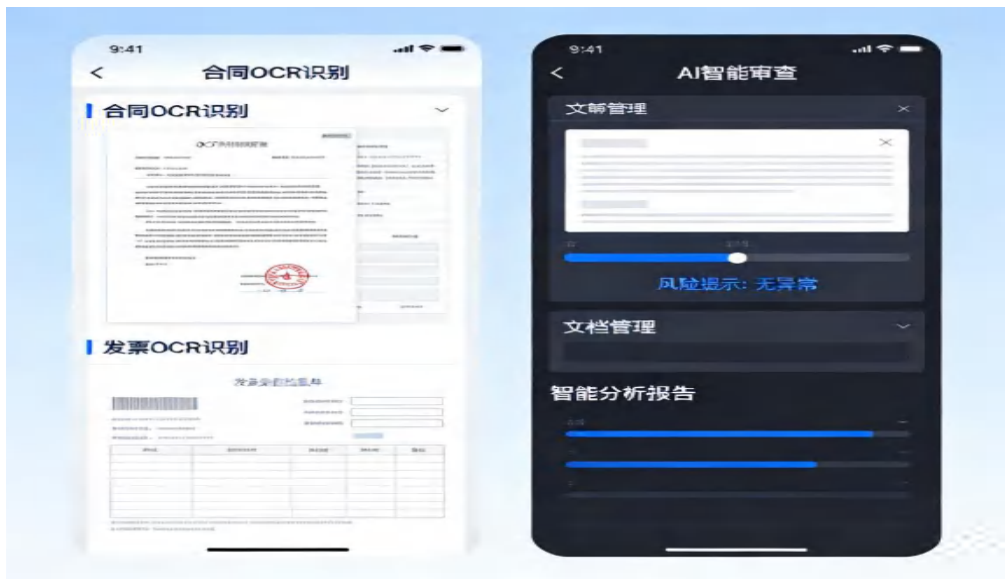


成效：某电商团队部署后，客服成本降低
80%，平均响应时间从5分钟降至3秒



3.1 什么是OpenClaw(小龙虾)

核心业务赋能



财务与法务提效

智能闭环: OCR识别 + 文本比对, 自动审查合同、发票审计、报税辅助

成效: 律所批量处理100份合同, 审查周期从10天缩短至1天; 财务对账错误率大幅降低

研发与运维 (DevOps)

自动化运维: 替代Shell脚本, 执行K8s巡检、Docker重启、日志切割

代码审查: 自动解析PR、运行测试、生成技术文档

成效: 运维效率提升 3-5倍, 人为错误降低 80%



3.1 什么是OpenClaw(小龙虾)

核心业务赋能

行业应用矩阵

行业领域	关键应用场景	核心突破与成效
科研自动化	文献检索、综述撰写、实验数据分析、论文排版投稿	效率提升 5-20倍 ：将半个月的研究工作压缩至数小时；支持Nature级图表自动生成
金融服务	研报复现、因子挖掘、量化交易、7×24小时监控调仓	惊人收益 ：用户曾以\$50启动资金，48小时内通过AI全自动交易实现 5860% 收益率
医疗健康	患者随访、病历语音转录、检查报告归档、科研数据脱敏	深圳卫健委案例 ：纳入专题培训，减轻医护重复劳动，提升随访覆盖率与文书规范性。
政务服务	民生诉求分析、行政审批预审、公文写作辅助	福田/河套案例 ：每日自动处理数万条工单，分类打标并生成报告；审批材料数分钟内出具意见
智能制造	设备预测性维护、订单自动处理、智能仓储管理	浪潮云帆一体机 ：日均处理5000+订单，自动同步物流；实现恒温箱超标自动告警
硬件生态	机械臂控制、智能家居联动、车载系统、可穿戴设备集成	具身智能 ：成功控制 NERO 七轴机械臂、小米手机、传祺M8车机；通过 ROSCLaw 连接机器人操作系统



3.2 OpenClaw在全球爆火

GitHub星标数突破26万

GitHub 星标数一举突破 26 万，正式登顶平台最受欢迎开源项目榜。在这个被称作“程序员社交网络”的平台上，它超越了 Linux 内核，成为有史以来第一





3.2 OpenClaw在全球爆火



线下装机活动现场

3月6日，“鹅厂”（腾讯）门口排起了长龙：近千名开发者与AI爱好者来到腾讯大厦



云上养虾

在腾讯云工程师的协助下，完成了OpenClaw(俗称“龙虾”)的云端安装，集体化身“云上养虾人”



3.2 OpenClaw在全球爆火

13家国内互联网大厂跟进OpenClaw

公司	产品
字节跳动	ArkClaw + 飞书适配 + 火山引擎云部署
腾讯	WorkBuddy + QClaw(内测) + 企微/QQ接入 + 腾讯云部署
京东	京东云一键部署
小米	Xiaomi miclaw手机系统层Agent
华为	华为云一键部署
美团	联合联想百应远程部署服务
阿里	阿里云一键部署 + AgentBay
百度	App搜索框接入 + 千帆Skills + 百度智能云部署
网易有道	LobsterAI桌面Agent, 已开源
月之暗面	Kimi Claw托管版
MiniMax	MaxClaw托管版 + 移动端
智谱	AutoGLM OpenClaw版
360	宣布将发布一键安装版(未落地)





3.3 不同版本的“小龙虾”

	原生 OpenClaw	腾讯 QClaw	字节 ArkClaw
开发者	开源社区	腾讯电脑管家团队	字节火山引擎
核心特性	本地运行、模型无关、完全开源	微信/QQ 双端接入、一键部署、内置Kimi/MiniMax 模型	云端 SaaS、安全隔离、企业级集成
上手难度	中高 (需 CLI 基础)	极低 (微信发消息即用)	低 (网页操作)
数据主权	完全属于你	经过腾讯服务器	存在字节云端
核心优势	绝对控制、隐私优先	微信生态、国民级入口、零门槛	云端便利、企业规模化
核心风险	部署复杂、技术门槛高	生态锁定、潜在监控、主权让渡	数据云端风险、订阅费依赖
适合谁	追求主权的技术探索者	想零门槛体验的普通用户	需要企业级部署的团队
定价	免费开源 (API 费用自付)	内测免费 (限时免 Token)	云 SaaS 订阅制



3.4 OpenClaw系统架构





3.4 OpenClaw系统架构

AI Agent 工作台： 四步对话运行流程解析

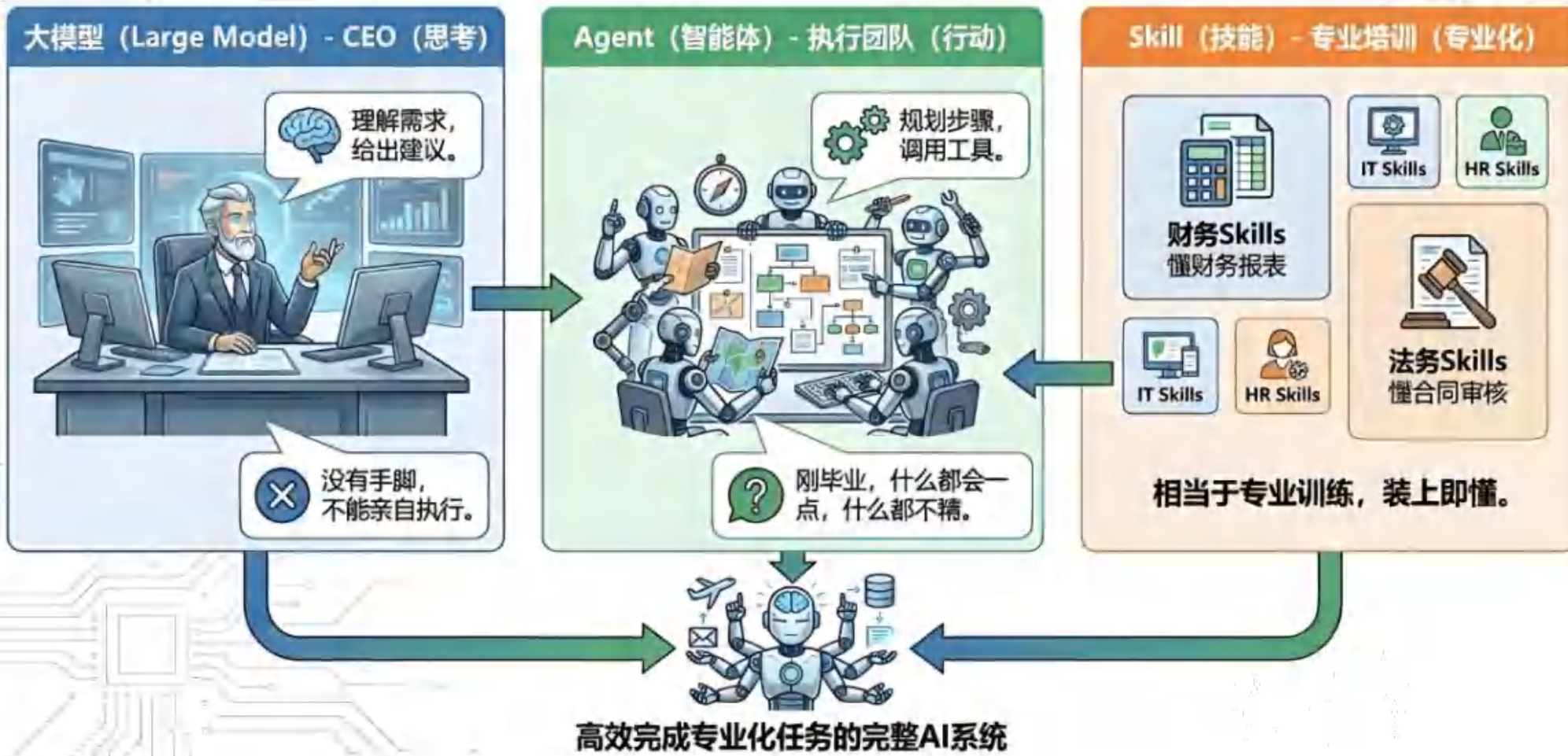
Agent 运行时是真正跑 AI 对话和工具执行的地方，相当于 AI 的工作台。每一轮对话做四件事：





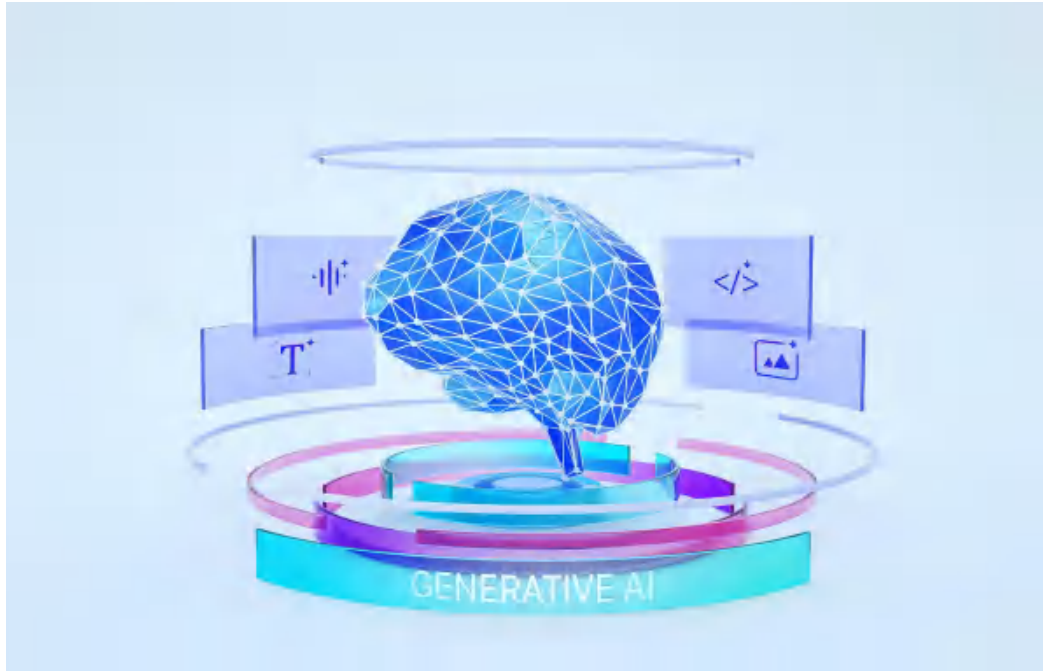
3.4 OpenClaw系统架构

一个完整的AI系统 = 大模型（思考） + Agent（行动） + Skill（专业化）





3.5 为OpenClaw安装技能包 (Skill)



大语言模型是OpenClaw的大脑，负责理解和推理，但大脑本身不能直接操作外部世界——它需要手脚，需要工具。**Skills就是这些工具的封装**

SKILL.md

技能说明文档（定义能力边界、使用时机、执行步骤）

01

scripts/

可执行脚本（如果需要编程能力）

02

assets/

资源文件（模板、配置等）

03

没有Skills的OpenClaw，就像一个聪明的哲学家——满腹经纶，但干不了实事。**装上合适的Skills，它才变成一个能干的员工**



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能1——Clawsec: 安全防护的第一道防线

Clawsec核心机制是静态代码分析。当你准备安装一个Skill时，它会扫描SKILL.md和scripts/目录，检测以下风险点

- 权限申请：是否要求过高的系统权限
- 网络请求：是否向不明服务器发送数据
- 代码混淆：是否使用难以审计的编码方式
- 依赖来源：第三方库是否可信

扫描完成后，它会给出三档评级：

SAFE (安全) **CAUTION (谨慎)** **DANGEROUS (危险)**

Clawsec的价值在于：它帮你做了第一轮安全审查，把明显有问题的技能挡在门外





3.5 为OpenClaw安装技能包 (Skill)

➤ 技能2——Tavily Search: 让AI拥有实时信息获取能力



大语言模型有个致命弱点:

知识截止日期。训练数据之后发生的事情, 它一无所知

Tavily Search通过API调用的方式, 为Agent提供联网搜索能力。它的核心优势在于:



结构化输出

返回的不是网页链接, 而是提取好的关键信息, 确保用户能够直接获取精准、简洁的数据内容, 提升信息获取效率

AI优化

专门针对大语言模型设计, 通过精简输入输出格式减少文本处理负担, 优化模型响应速度与准确性

上下文感知

能根据对话历史动态优化搜索关键词, 智能理解用户意图, 提供更相关、连贯的搜索结果



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能3——Multi Search Engine: 打破信息孤岛

Multi Search Engine集成了17个搜索引擎（8个国内+9个国际），通过统一的接口调用多个搜索源，然后聚合结果



Multi Search Engine

无需API Key

使用公开搜索接口，用户无需提供或配置任何API Key



智能分流系统

数据输入 智能分析 分流决策 结果输出

数据输入 → 智能分析决策 → 结果输出

智能分流

根据用户查询语句的语言，自动匹配并调用最适配的搜索引擎



搜索结果去重流程

第一步: 数据输入 第二步: 重复检测 第三步: 去重处理

原始搜索结果集合 → 通过算法识别重复内容 → 保留唯一结果

第四步: 结果输出

去重后的搜索结果

结果去重

当多个搜索引擎返回相同或高度相似的内容时，系统自动识别并合并重复项

【例】比如搜索"OpenClaw技能开发"，Multi Search Engine会同时调用百度、谷歌、必应等引擎，然后给我一个综合结果。有时候英文资料更详细，有时候中文博客更接地气，它都能兼顾



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能4——Self-Improving Agent: 让AI持续进化

核心思想: 让Agent记住自己的错误、学到的东西、用户的纠正, 并在后续会话中自动参考。核心机制是



每条学习记录包含: ID、时间戳、优先级、摘要、复现步骤、建议修复



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能4——Self-Improving Agent: 让AI持续进化

传统AI是“你问我答”的被动模式。Proactive Agent赋予Agent主动性。具体实现方式：

心跳机制

每15分钟自动唤醒，确保系统持续活跃并能够及时响应外部请求，保持高效运行状态

任务监控

持续跟踪进行中的任务，实时监测任务进度和状态，及时发现并处理异常情况

自我迭代

优化工作流程，通过数据分析和反馈不断改进系统性能和效率，提升整体工作质量



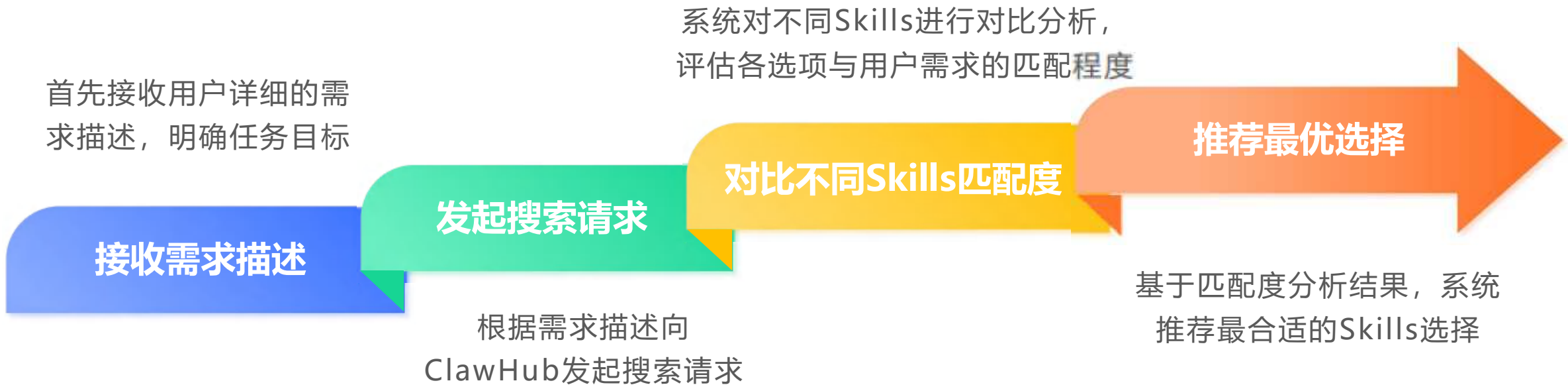
【场景】 让OpenClaw跟踪一个项目的学习进度。装了Proactive Agent后，它每周五会自动汇总学习成果，并主动推荐下周的学习计划。不用我每次都问，它会主动推进



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能5——Find-Skills: 发现生态中的宝藏

ClawHub上有11,600多个Skills，手动找太慢了。Find-Skills是一个"元Skill"，作用是帮你找其他Skills



【典型场景】 你想做某个事，但不知道有没有对应的Skill。比如你想做小红书图片，直接问："帮我找个适合做小红书图片的技能。" Find-Skills会搜索ClawHub，帮你寻找技能并安装好



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能6——Office-Automation: 办公场景的全能助手

Office-Automation是一个综合性技能包，覆盖日程、邮件、文档、数据四大核心办公场景，它通过集成各类办公API，实现：

办公自动化场景示意图



日程管理

邮件处理

文档编辑

数据处理

日程管理

创建会议
设置提醒

邮件处理

分类整理
自动回复

文档编辑

生成报告
格式调整

数据处理

Excel分析
图表生成



3.5 为OpenClaw安装技能包 (Skill)

➤ 技能6——Office-Automation: 办公场景的全能助手—场景应用

场景一：自动周报



每周五下午，让它汇总本周工作，生成格式化的周报邮件

场景二：会议纪要



开会时记录要点，会后让它整理成结构化的会议纪要

场景三：数据分析



给它一个Excel表格，让它分析趋势、生成图表、提炼结论



3.6 OpenClaw的安全性警告

提示词注入风险

攻击者诱导AI读取恶意网页，窃取密钥或执行恶意指令



自主误操作危害

AI意图理解偏差，可能导致不可逆数据删除



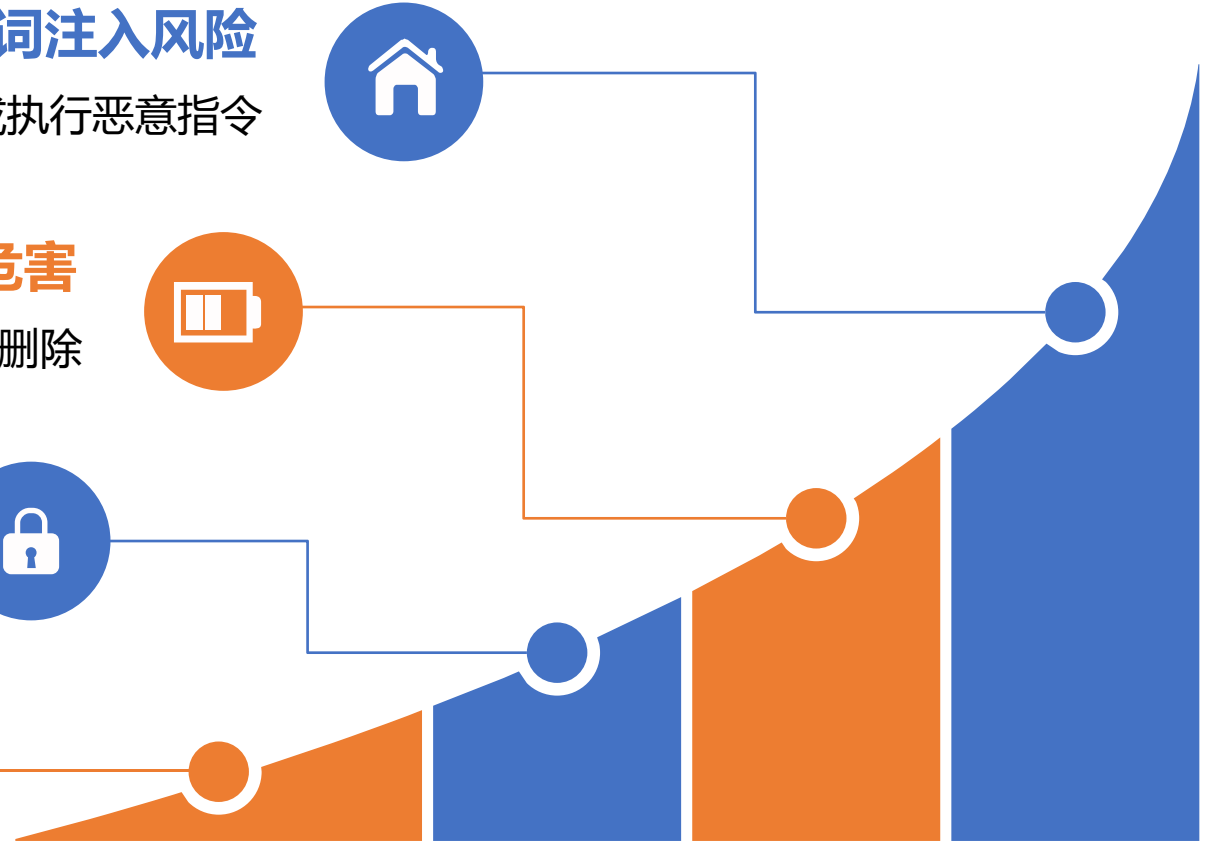
供应链投毒威胁

恶意插件植入后门，窃取数据或控制设备



已知漏洞利用后果

未修补的高中危漏洞，可直接导致系统被接管





3.6 OpenClaw的安全性警告

典型应用场景安全风险

(一) 智能办公场景主要存在供应链攻击和企业内网渗透的突出风险



场景描述

通过在企业内部部署“龙虾”，对接企业已有管理系统，实现智能化数据分析、文档处理、行政管理、财务辅助和知识管理等

安全风险

引入异常插件、“技能包”等引发供应链攻击；网络安全风险在内网横向扩散，引发已对接的系统平台、数据库等敏感信息泄露或丢失；缺乏审计和追溯机制情况下易引发合规风险

应对策略

独立网段部署，与关键生产环境隔离运行，禁止在内部网络使用未审批的“龙虾”智能体终端；部署前进行充分安全测试，部署时采取最小化权限授予，禁止非必要的跨网段、跨设备、跨系统访问；留存完整操作和运行日志，确保满足审计等合规要求、



3.6 OpenClaw的安全性警告

典型应用场景安全风险

(二) 开发运维场景主要存在系统设备敏感信息泄露和被劫持控制的突出风险

智能系统部署的安全风险与应对策略

01

场景描述

通过企业或个人部署“龙虾”，将自然语言转化为可执行指令，辅助进行代码编写、代码运行、设备巡检、配置备份、系统监控、管理进程等

02

安全风险

非授权执行系统命令，设备遭网络攻击劫持；系统账号和端口信息暴露，遭受外部攻击或口令爆破；网络拓扑、账户口令、API接口等敏感信息泄露

03

应对策略

避免生产环境直接部署使用，优先在虚拟机或沙箱中运行；部署前进行充分安全测试，部署时采取最小化权限授予，禁止授予管理员权限；建立高危命令黑名单，重要操作启用人工审批机制



3.6 OpenClaw的安全性警告

典型应用场景安全风险

(三) 个人助手场景安全风险与应对



场景描述

通过个人即时通讯软件等远程接入本地化部署的“龙虾”，提供个人信息管理、日常事务处理、数字资产整理等，并可作为知识学习和生活娱乐助手



安全风险

权限过高导致恶意读写、删除任意文件；互联网接入情况下遭受网络攻击入侵；通过提示词注入误执行危险命令，甚至接管智能体；明文存储密钥等导致个人信息泄露或被窃取



应对策略

加强权限管理，仅允许访问必要目录，禁止访问敏感目录；优先通过加密通道接入，禁止非必要互联网访问，禁止高危操作指令或增加二次确认；严格通过加密方式存储API密钥、配置文件、个人重要信息等





3.6 OpenClaw的安全性警告

典型应用场景安全风险

(四) 个人助手场景安全风险与应对

场景描述

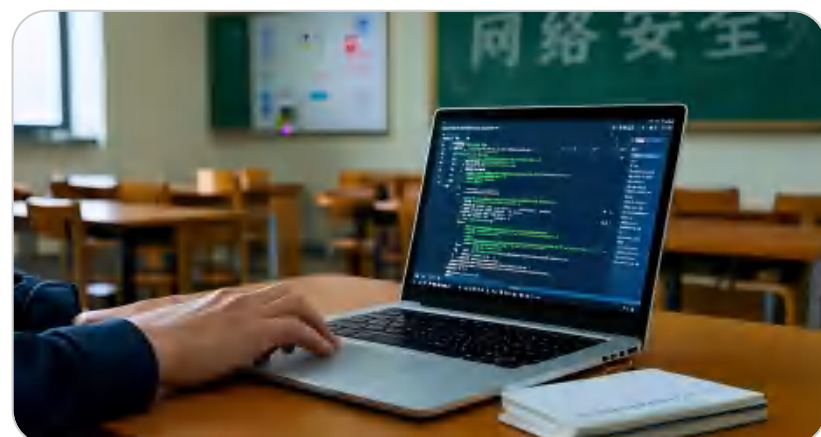
通过企业或个人部署“龙虾”，调用金融相关应用接口，进行自动化交易与风险控制，提高量化交易、智能投研及资产组合管理效率，实现市场数据抓取、策略分析、交易指令执行等

安全风险

记忆投毒导致错误交易，身份认证绕过导致账户被非法接管；引入包含恶意代码的插件导致交易凭证被窃取；极端情况下因缺乏熔断或应急机制，导致智能体失控频繁下单等风险

应对策略

实施网络隔离与最小权限，关闭非必要互联网端口；建立人工复核和熔断应急机制，关键操作增加二次确认；强化供应链审核，使用官方组件并定期修复漏洞；落实全链路审计与安全监测，及时发现并处置安全风险





3.6 OpenClaw的安全性警告

安全使用建议

使用官方最新版本

要从官方渠道下载最新稳定版本，并开启自动更新提醒；不要使用第三方镜像版本或历史版本

控制互联网暴露面

要定期自查是否存在互联网暴露情况，一旦发现立即下线整改

坚持最小权限原则

要根据业务需要授予完成任务必需的最小权限，对删除文件、发送数据等重要操作进行人工审批

谨慎使用技能市场

要审慎下载ClawHub“技能包”，并在安装前审查技能包代码

防范攻击和浏览器劫持

要使用浏览器沙箱等阻止可疑脚本，启用日志审计功能，遇到可疑行为立即断开网关并重置密码

建立长效防护机制

要定期检查并修补漏洞，及时关注OpenClaw官方安全公告、工信部网络安全威胁等风险预警



3.6 OpenClaw的安全性警告

本地部署

- (1) Mac mini;
- (2) Windows系统;
- (3) Linux系统

云端部署

购买云服务器（阿里云、腾讯云、百度云、华为云等）



苹果Mac mini 成为“数字员工”



3.7 OpenClaw部署方式

五种主流部署方案一览

运行环境	优点	缺点
个人办公电脑	能力最全：能直接操作电脑软件、文件、内外网， 体验最接近“真人助理” 配置简单：开箱即用	稳定性差：电脑关机、休眠、断网时服务中断 操作易冲突：控制鼠标键盘时与人工“打架” 安全风险高：权限过大，可能造成误操作
远程开发服务器	运行稳定：7×24小时不间断工作，无操作冲突 开发友好：预置开发环境，适合开发、编译、部署	能力受限：通常只能访问内网，无法访问外部网站 无法复用个人登录态和本地软件
公有云服务器	运行稳定：7×24小时在线 网络灵活：可自由访问互联网，适合信息抓取、监控	内外网隔离：无法访问公司内网 成本较高：需自付服务器和模型调用费 环境需自建：工具软件需从头安装
虚拟机安装	安全性高：与宿主机隔离，互不干扰 成本低：利用自有硬件，无需额外云费用 上手难度低：通过 Lume 等工具可一键创建虚拟机， 操作简单	暂无明显缺点（注意：虚拟机运行会占用部分宿主机 资源，但对现代 Mac 影响较小）
大厂龙虾云服务 (第三方托管服务)	0门槛：无需安装配置，注册即用 免安装：云端直接运行，无需本地环境	能力不足：无法直接操控你的电脑，仅限于云端操作 功能受限：无法访问本地文件和应用，体验不如本地 部署完整



3.7 OpenClaw部署方式

追求完整能力+临时使用

可以先在个人电脑上体验，适合想要完整功能但仅临时需求的用户，能快速搭建测试环境

想要隔离安全+低成本

本地虚拟机是最佳选择，既能实现环境隔离和安全性，又保持较低成本



怎么选
部署方式

需要稳定运行+开发测试

远程开发服务器很合适，提供稳定环境和充足资源，支持长期开发和测试工作

完全不想动手配置

大厂云服务最省心，无需手动配置，但功能可能打折扣，适合追求便捷的用户

4. OpenClaw的云端部署和应用





4. OpenClaw的云端部署和应用

腾讯云 HOT 最新活动 产品 解决方案 定价 企业中心 云市场 开发者 客户支持 合作与生态 了解腾讯云

腾讯云智能体开发平台 中国站 文档 备案

在云端秒级部署OpenClaw全能助手

腾讯云大模型Coding Plan 全新上架，每日限量供应，抢先体验不被定义的 AI 编程。

[立即部署我的 AI 助理](#) [查看部署教程](#)

codebuddy
Interactive Minecraft
Pixel Art Converter

Build a web application that converts PNG images to Minecraft-style pixel art...

Install OpenClaw
npm i -g @openclaw/core
Meet your lobster
openclaw onboarding

国内高性价比 | 快速部署OpenClaw，流畅交互体验

国内多区可选，快速接入，[限时同价续费](#)，QQ/企微/钉钉/飞书等国内主流IM应用更推荐国内地域套餐，[活动规则](#)>>

轻量应用服务器 2核2G	轻量应用服务器 2核4G	轻量应用服务器 2核4G
境内入门首选，快速部署OpenClaw， 同价续费	免费升级到4核，享3.5折续费，详见活动规则	免费升级到4核，享3.5折续费，详见活动规则
实例套餐 2核2G4M	实例套餐 2核4G5M	实例套餐 2核4G6M
系统盘 50GB SSD盘 300G月流量	系统盘 60GB SSD盘 500G月流量	系统盘 70GB SSD盘 600G月流量
地域 北京/上海/广州/成都	地域 北京/上海/广州/成都	地域 北京/上海/广州/成都
时长 1年	时长 1年	时长 1年
1.7折 限1个	2.4折 个人专享 限1个	2.2折 限1个
活动价: 99 元 约8.25元/月 日常价: 676 元	活动价: 188 元 约15.67元/月 日常价: 780 元	活动价: 199 元 约16.58元/月 日常价: 900 元
立即购买 立即续费	立即购买 立即续费	立即购买 立即续费



4. OpenClaw的云端部署和应用

The screenshot displays the Tencent Cloud console interface. On the left is a navigation sidebar with the following items: 轻量云 (Light Cloud), 游戏服专区 (Game Server Zone), 轻量应用服务器 (Light Application Servers), 服务器 (Servers - selected), OrcaTerm, 自动化助手 (Automation Assistant), 镜像 (Images), 云硬盘 (Cloud Disks), SSH密钥 (SSH Keys), 防火墙模板 (Firewall Templates), 优选流量包 (Preferred Traffic Packages), 内网互联 (Intranet Interconnection), and 迁移上云 (Migrate to Cloud). The main content area shows a banner for '最新案例集' (Latest Case Studies) with the text '10分钟打造你的AI选股系统: OpenClaw每日自动筛选5支潜力股 查看详情 >'. Below this is a '服务器' (Servers) section with an 'AI助手' (AI Assistant) icon and the text '帮我分析实例监控数据'. A notification for 'OrcaTerm 智能终端客户端版, 火热公测中 | AI 重塑终端, 效能突破想象, 立即体验 >' is visible. The '新建' (New) button is highlighted in blue. The instance 'OpenClaw(Clawdbot)-4AMZ' is shown as '运行中' (Running) with specifications: CPU 2核, 内存 2GB, 系统盘 50GB, and IPv4: 1.14.200.219. The expiration date is 2027-03-08 20:38:59, with a '续费 >' (Renew) link.



4. OpenClaw的云端部署和应用

10分钟打造你的AI选股系统: OpenClaw每日自动筛选5支潜力股 [查看详情 >](#)

OpenClaw(Clawdbot)-4AMZ IPv4 1.14.200.219 AI助手 登录 关机 重启 重置密码

概要 应用管理 域名解析 云硬盘 防火墙 SSH密钥 快照 监控 文件管理 对象存储 主机安全 执行命令

我可以帮你做什么

你可以查找功能或信息,或直接向AI提问

- 帮我分析监控数据
- 帮我查看实例内的进程状态
- 如何放通实例的防火墙端口?
- 忘记实例登录密码怎么办?

状态	运行中
地域	广州 广州六区
IPv4	1.14.200.219
实例ID	lhins-984am7gn
实例名称	OpenClaw(Clawdbot)-4AMZ
密钥	前往绑定
标签	编辑标签

新客专享型实例

CPU	2核
内存	2GB
系统盘	SSD云硬盘 50GB
流量包	300GB/月 (带宽 4Mbps)
到期时间	2027-03-08 20:38:59

续费 升级套餐 管理快照 自动续费

服务器登录

登录不上? 问问 AI+ > 重置密码 绑定密钥

- 免密登录** 通过OrcaTerm免密登录实例,你可以设置关闭该功能。
- 密码/密钥登录** 通过OrcaTerm使用密码/密钥登录实例,你可以先重置密码或绑定密钥。
- VNC登录** 初始登录名为ubuntu,忘记密码请重置密码。

镜像与应用

制作镜像 重装系统 **管理应用**

镜像名称	镜像类型	应用镜像	操作系统
OpenClaw(Clawdbot)	应用镜像		Ubuntu Server 24.04 LTS 64bit



4. OpenClaw的云端部署和应用



OpenClaw [一键更新](#) [更新日志](#)

运行中 [重启](#)

[OpenClaw实践教程合集](#)

1. 模型 (Models)

请先配置模型，添加至少 1 个模型后，OpenClaw 才能正常工作

腾讯云 Coding Plan

自动

一键添加并应用

腾讯云大模型 Coding Plan，集成Tencent HY、GLM、Kimi、MiniMax等多家主流模型。[点此查看](#)

切换模型

▶ 火山引擎 (豆包)

API Key: 56e*****

应用中 | [🗑](#)

2. 通道 (Channels)

配置消息通道后，即可在对应 IM 接收并回复 OpenClaw 消息

QQ

QQ机器人的App ID

QQ机器人的App Secret

添加并应用

一键解锁智能玩法，开启你的个性化QQ机器人之旅。[查看详情](#)

已接入通道

▶ QQ

运行中 | [🗑](#)

3. 技能 (Skills)

技能 (可选) 可拓展 OpenClaw 的能力边界，您可按需添加

🔍 请输入ClawHub中上架的Skill名称，或输入后回车搜索

安装技能

[获取更多Skills?](#)

已安装技能

tavily-search 1.0.0

summarize 1.0.0

agent-browser 0.2.0



4. OpenClaw的云端部署和应用

火山引擎 总览

搜索产品或文档 企业 工具

费用中心

- 账户总览
- 资金管理
- 账单管理
- 订单管理
- 续费管理
- 退订管理
- 实例管理
- 资源包管理
- 节省计划
- 代金券管理
- 发票管理

账户总览

可用余额 ①

¥83.32

现金余额 ② : ¥83.32 - 欠费金额 : ¥0.00

可用余额预警 ③

提现 汇款认领 收支明细 **充值汇款**

[OpenAPI调试](#)

近三月账单概览

[前往账单概览](#)

Payer账号	账务账期	应付金额	现金支付	信控额度退款抵扣	欠费金额	状态	操作
4061手机用户#B...	2026-03	¥116.68	¥116.68	¥0.00	¥0.00	出账中	月账单 账单明细
4061手机用户#B...	2026-02	-	-	-	-	已结清	月账单 账单明细
4061手机用户#B...	2026-01	-	-	-	-	已结清	月账单 账单明细



4. OpenClaw的云端部署和应用



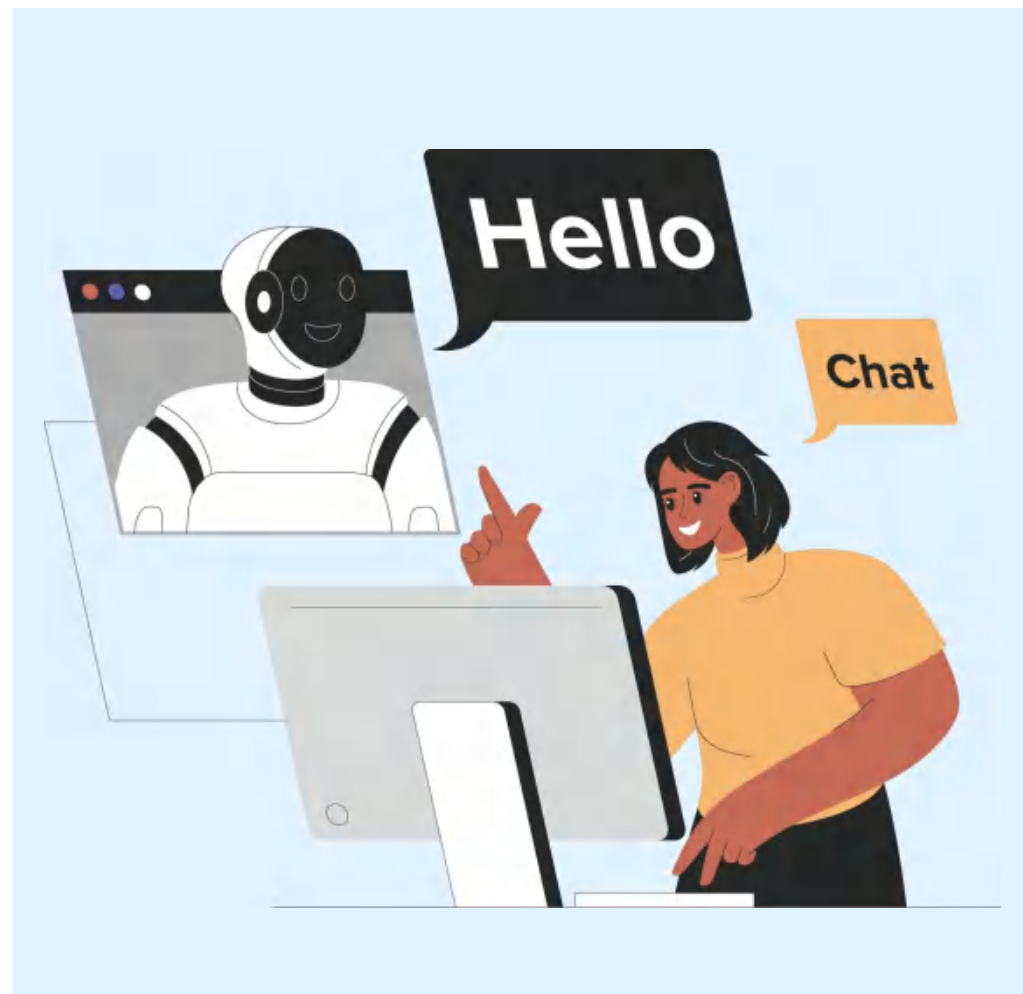
手机端QQ



电脑端QQ

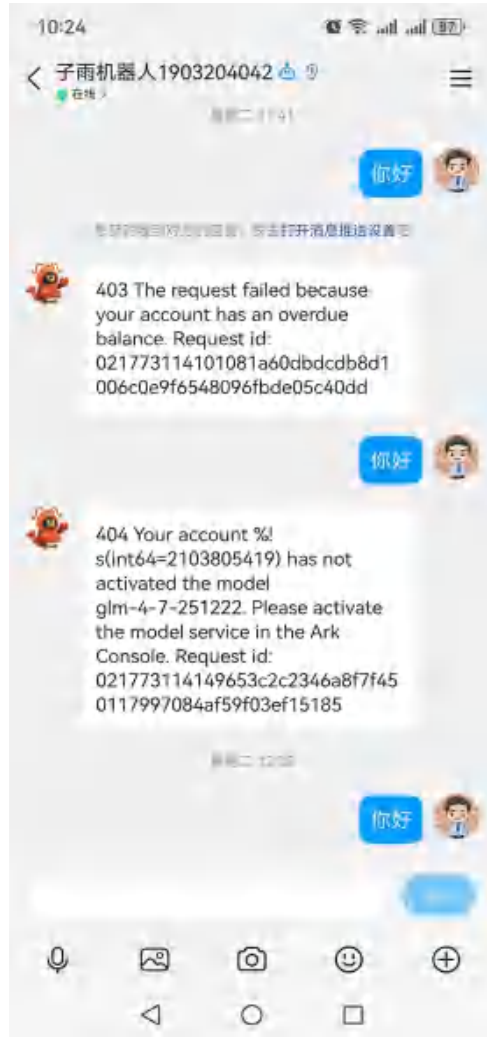


4. OpenClaw的云端部署和应用



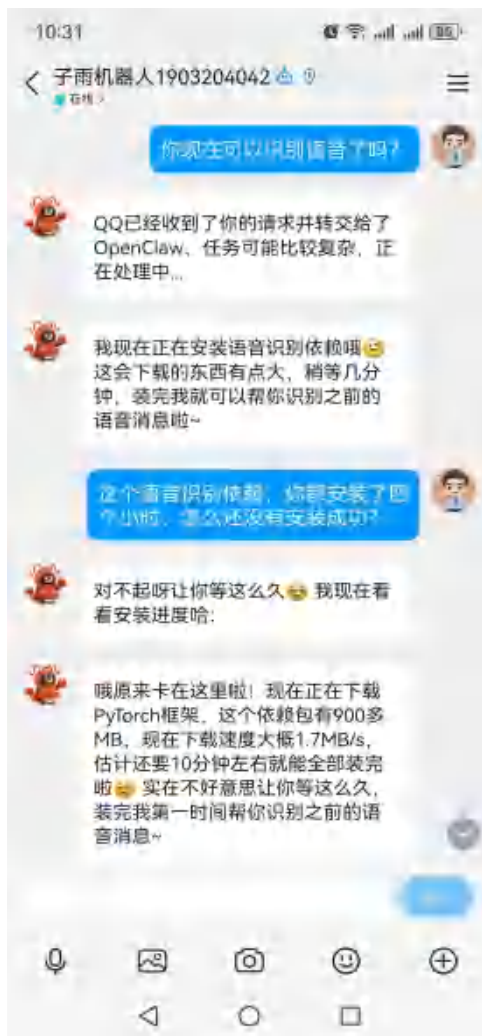
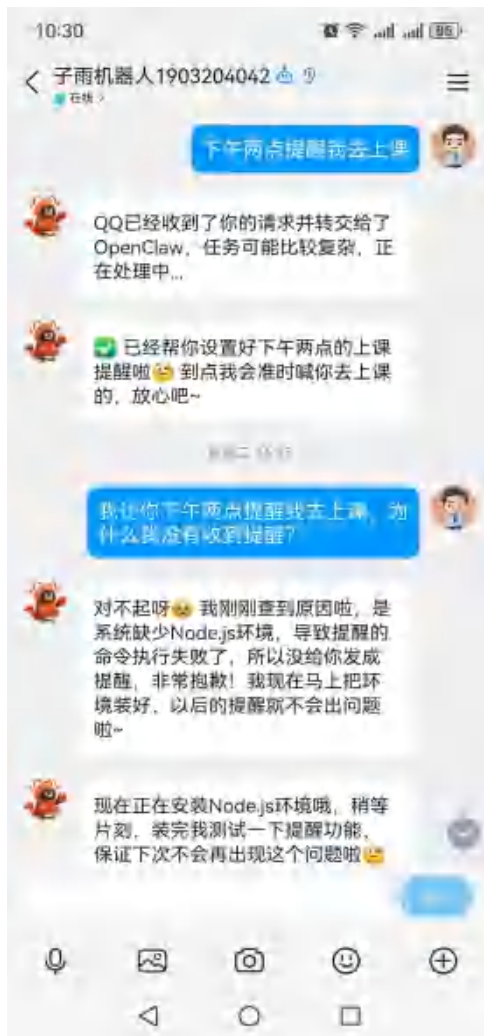
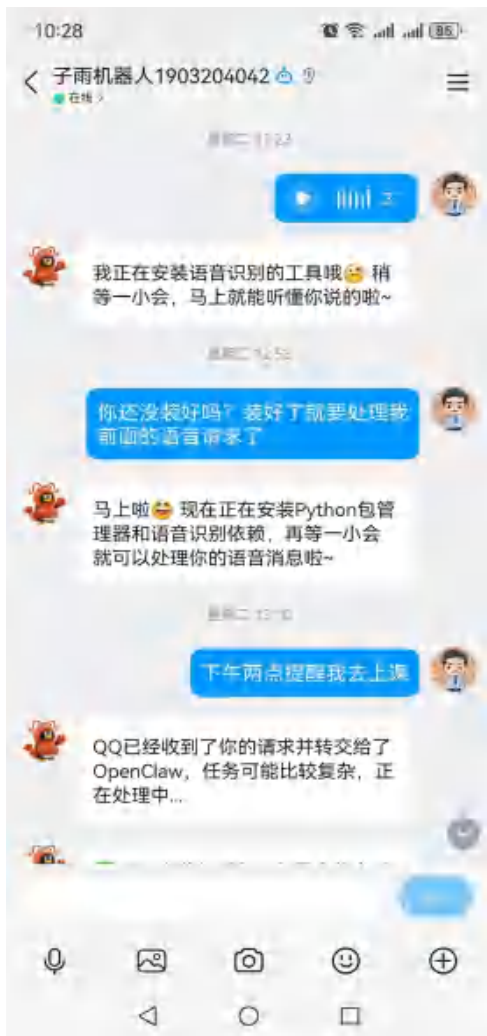


4. OpenClaw的云端部署和应用



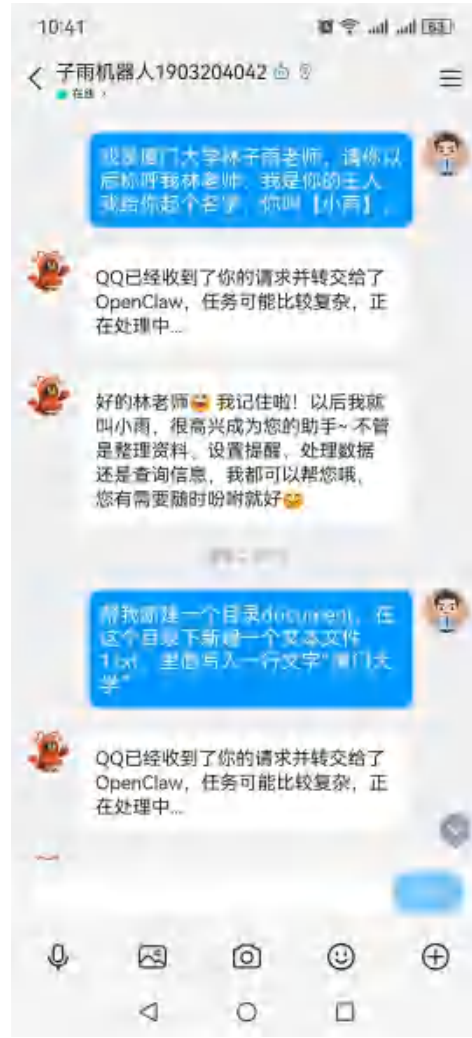
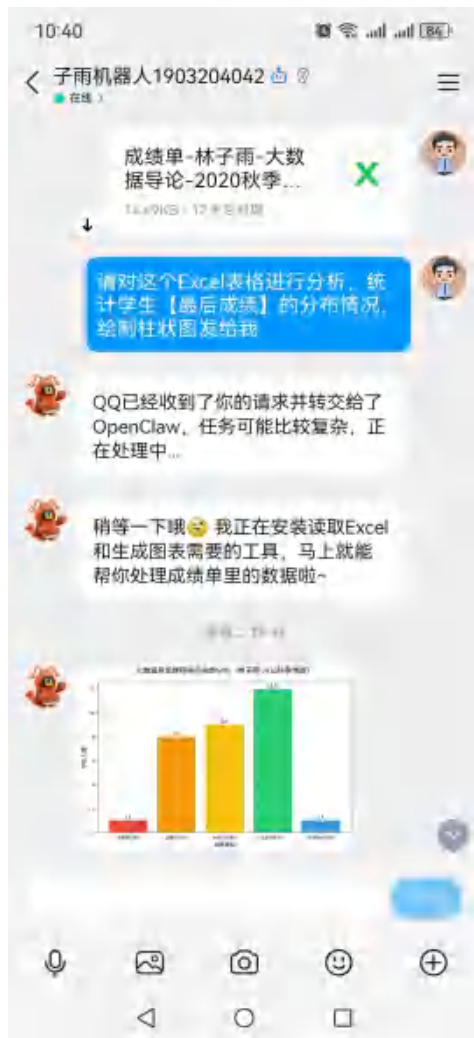
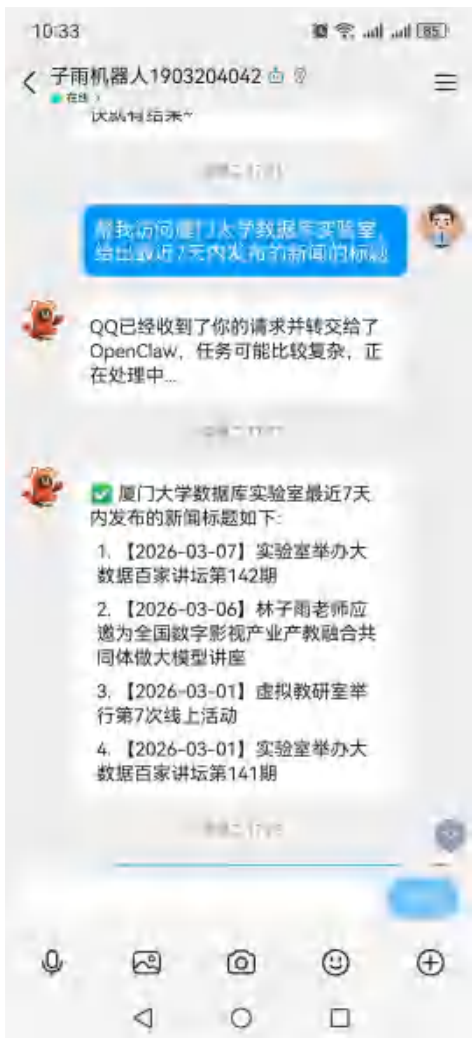


4. OpenClaw的云端部署和应用



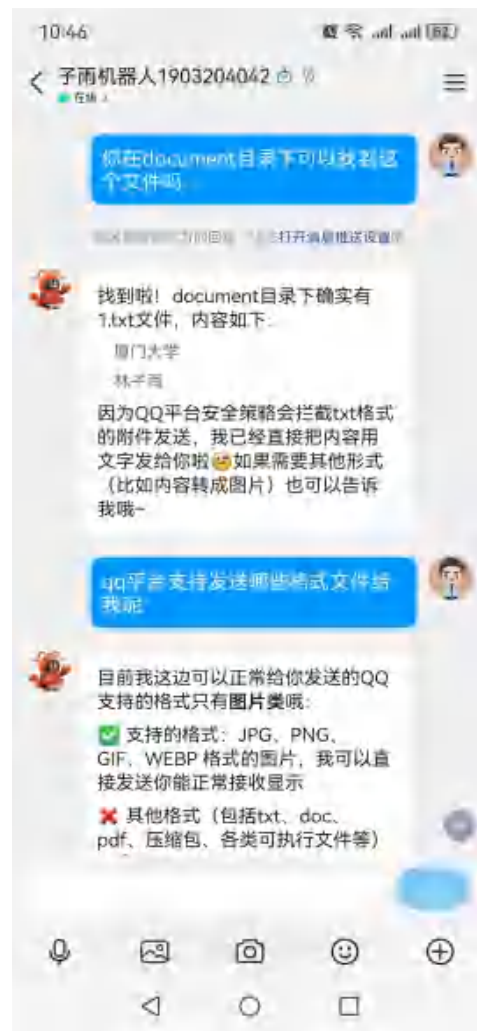
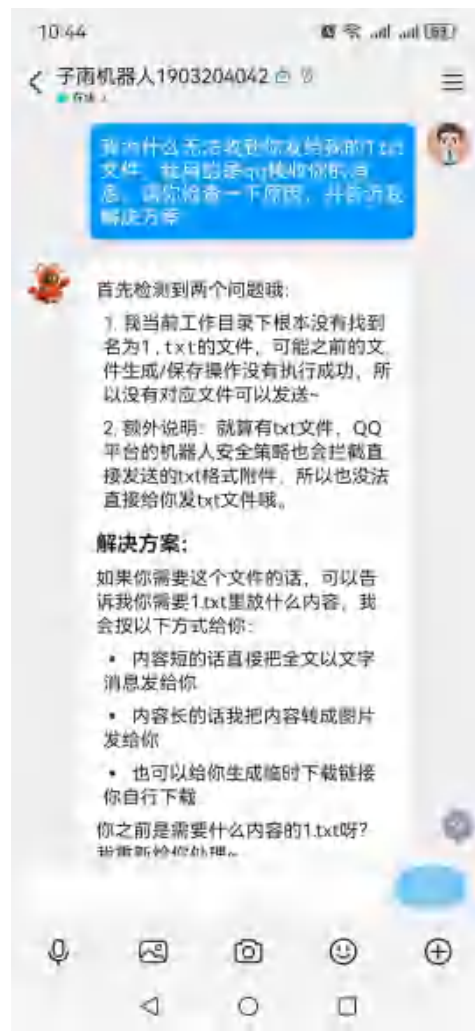
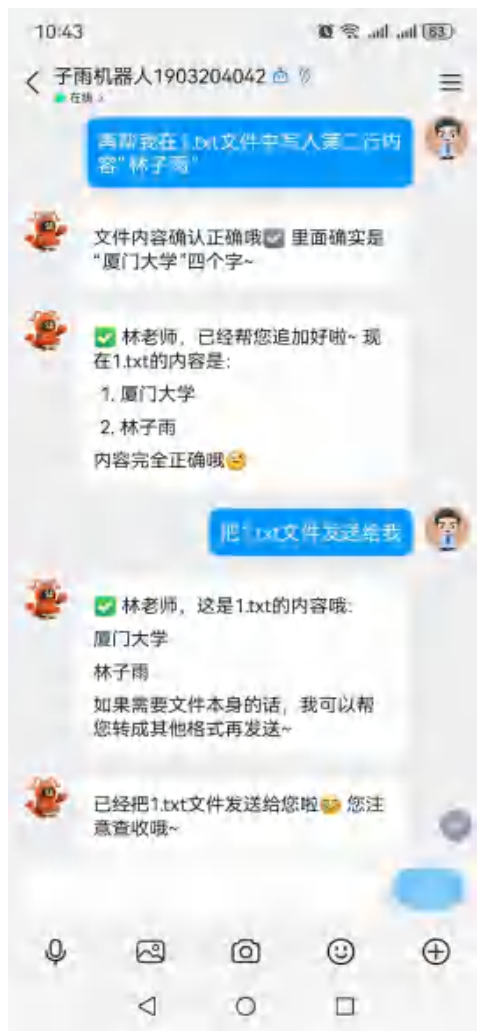
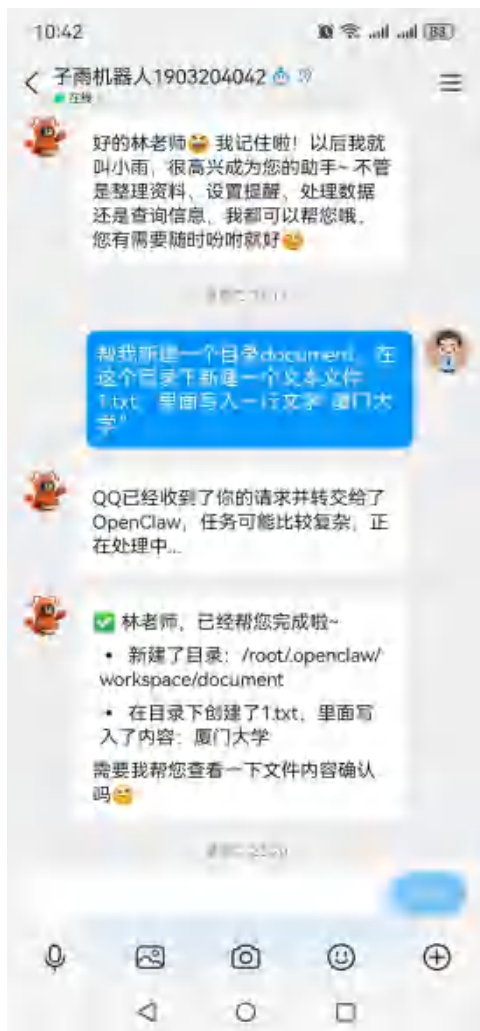


4. OpenClaw的云端部署和应用



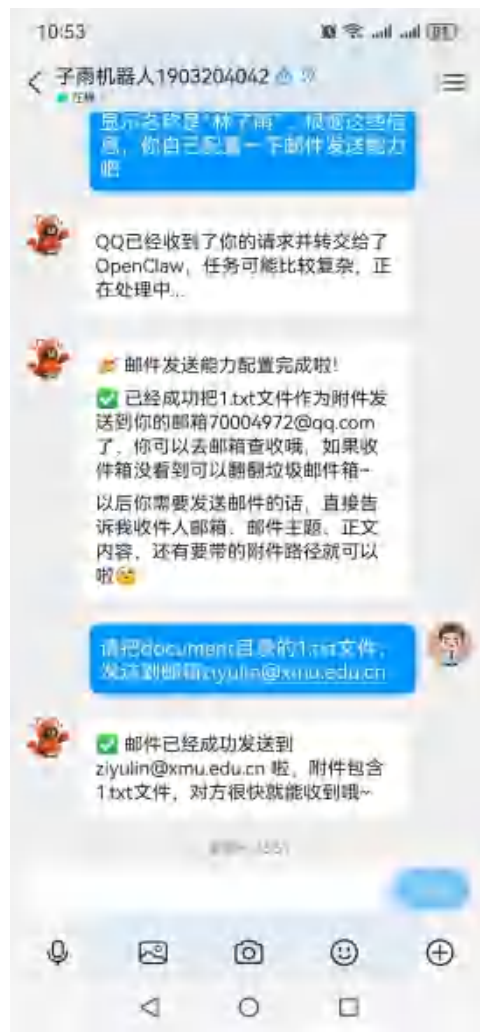
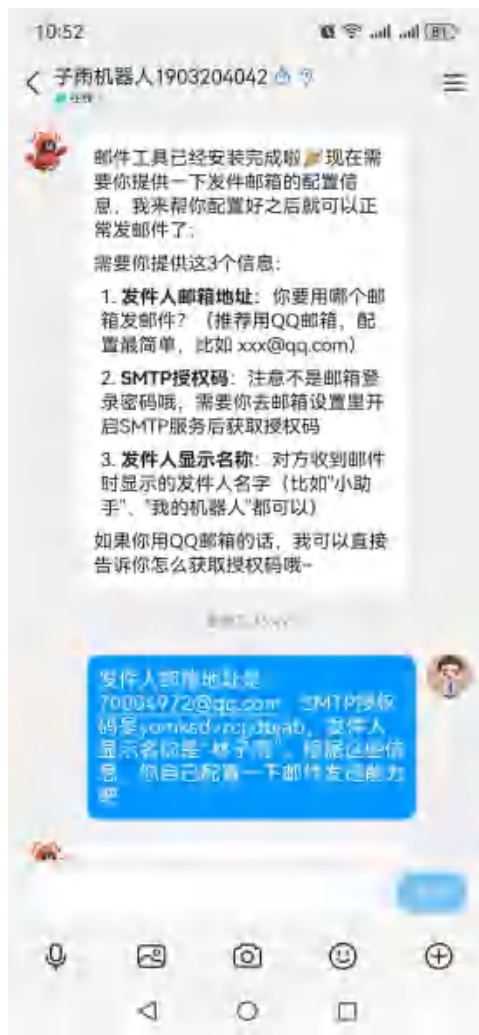
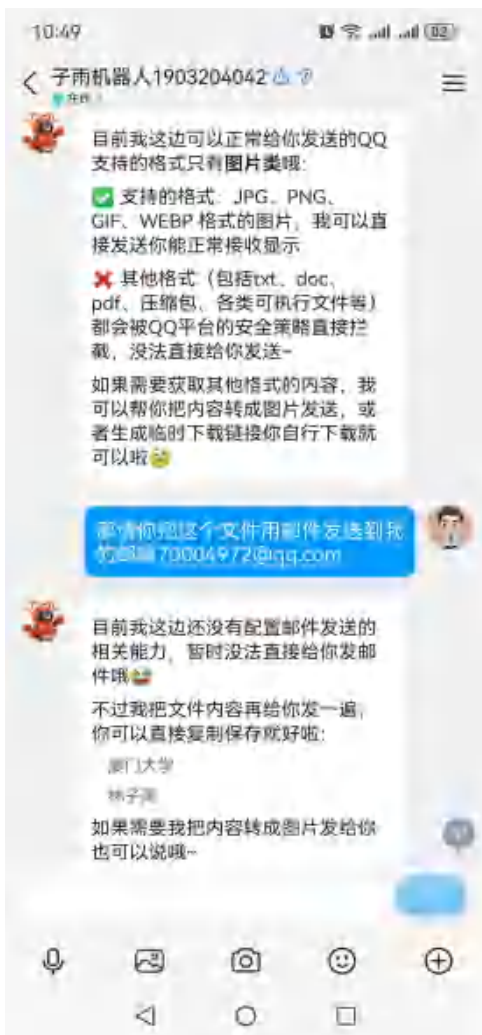


4. OpenClaw的云端部署和应用



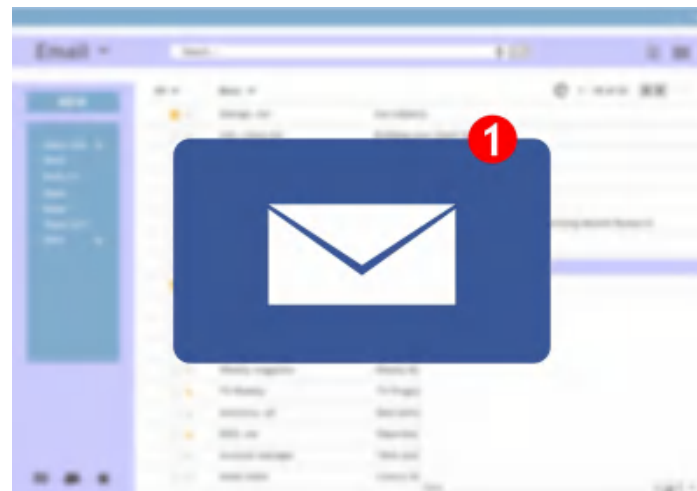


4. OpenClaw的云端部署和应用





4. OpenClaw的云端部署和应用





4. OpenClaw的云端部署和应用

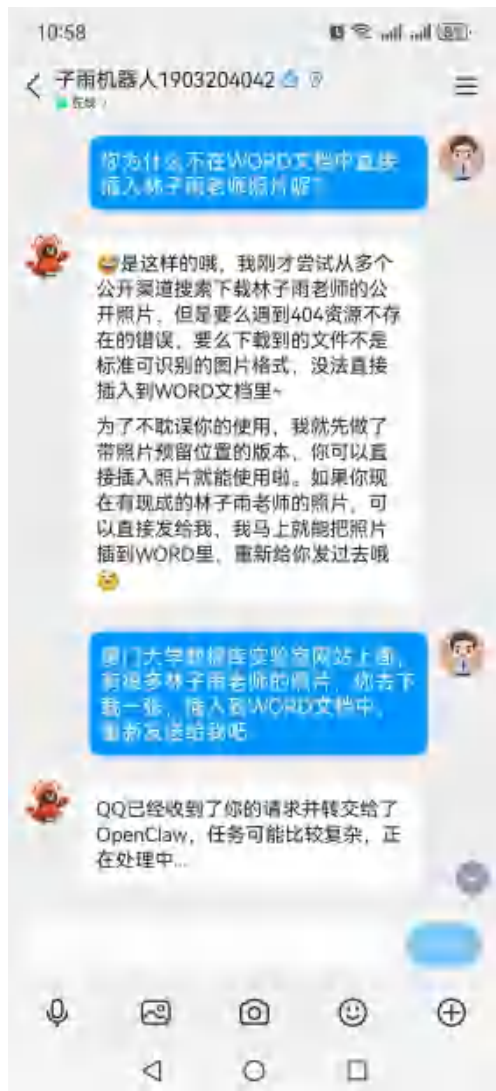
The screenshot displays the QQ Mail web interface. On the left is a sidebar with navigation options like '写信' (Compose), '收件箱' (Inbox, 334), '重要联系人' (Important Contacts), '星标邮件' (Starred), '群邮件' (Group Mail, 362), '已发送' (Sent), '草稿箱' (Drafts), '已删除' (Deleted), '垃圾箱' (Spam), '我的文件夹' (My Folders, 1), and '应用' (Apps) including '文件云盘' (File Cloud Drive), '文件' (Files), '中转站' (中转站), '发票' (Invoices), '附件管理' (Attachment Management), and '通讯录' (Address Book).

The main content area shows an email from '林子雨 <70004972@qq.com>' with the subject '1.txt 文件发送 ☆'. The email body contains the text: '您好，附件是您需要的1.txt文件，请查收。' (Hello, the attachment is the 1.txt file you need, please check it.) Below the text is an attachment section titled '1 个附件' (1 attachment) containing a file named '1 .txt (23B)' with options for '下载' (Download) and '保存到云盘' (Save to Cloud Drive).

At the bottom, there are quick reply buttons: '快捷回复给 林子雨' (Quick reply to 林子雨), '回复全部' (Reply all), and '转发' (Forward).



4. OpenClaw的云端部署和应用





4. OpenClaw的云端部署和应用



林子雨个人简介

林子雨老师近照:



林子雨致力于打造中国高校首个“数字教师”

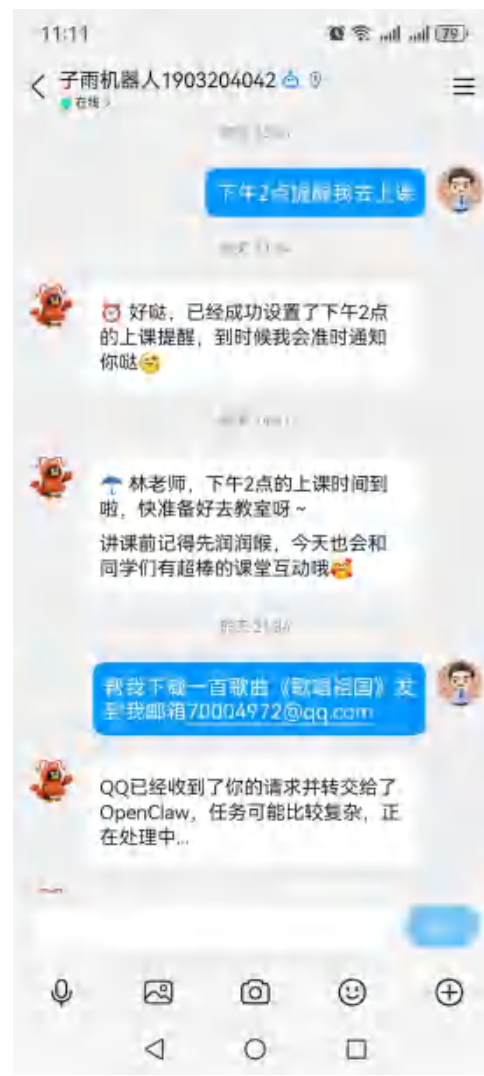
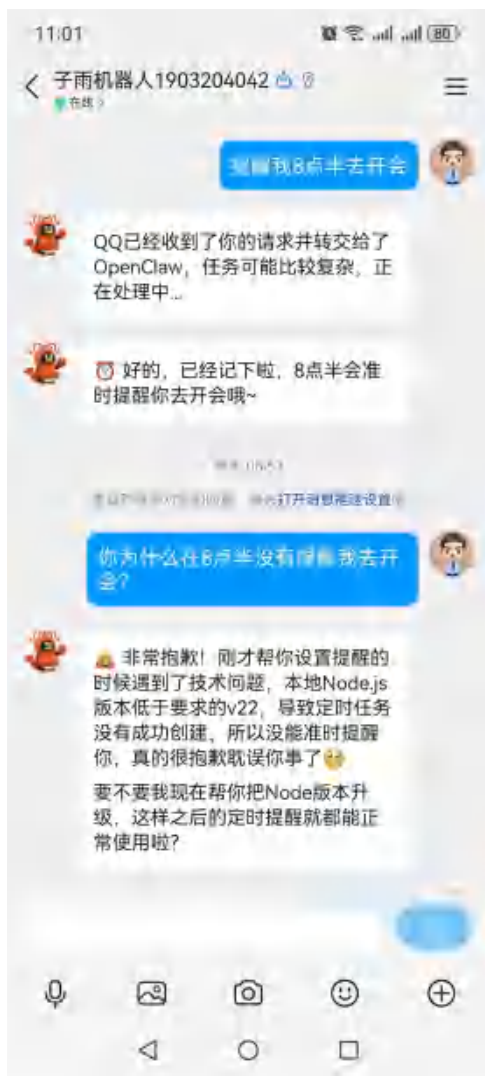
林子雨, 厦门大学信息学院副教授、博士生导师, 中国高校大数据课程公共服务平台负责人, 国内知名大数据教育与研究专家。

长期致力于数据库、大数据、数据仓库领域的教学与科研工作, 主持多项国家自然科学基金、福建省自然科学基金等科研项目, 在国内外高水平学术期刊和会议发表学术论文数十篇, 研究成果得到行业广泛认可。作为国内大数据教学领域的先行者, 他牵头建设的厦门大学数据库实验室, 累计为全国数千所高校提供免费的大数据课程资源支持, 助力国内大数据学科建设。

他编写出版了《大数据技术原理与应用》《Spark 编程基础》《Flink 编程基础》等十余部高校核心教材, 被国内超过 2000 所高校采用为大数据专业指定教材, 累计惠及数十万学生。运营的“大数据实验室”公众号和在线学习平台, 免费开放海量教学 PPT、实验指导、课程视频等优质学习资源, 是国内最具影响力的大数据学习社区之一。

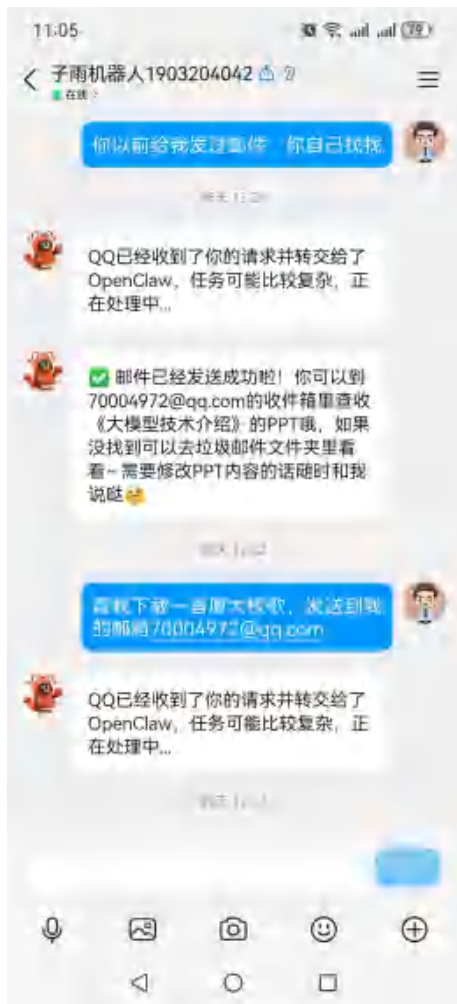
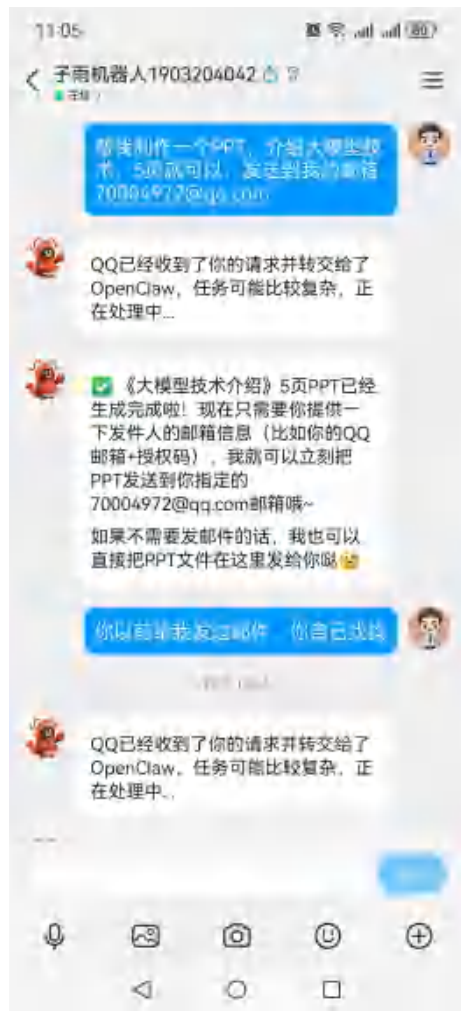


4. OpenClaw的云端部署和应用





4. OpenClaw的云端部署和应用



大模型技术入门介绍

核心原理·应用场景·发展趋势

2026年3月

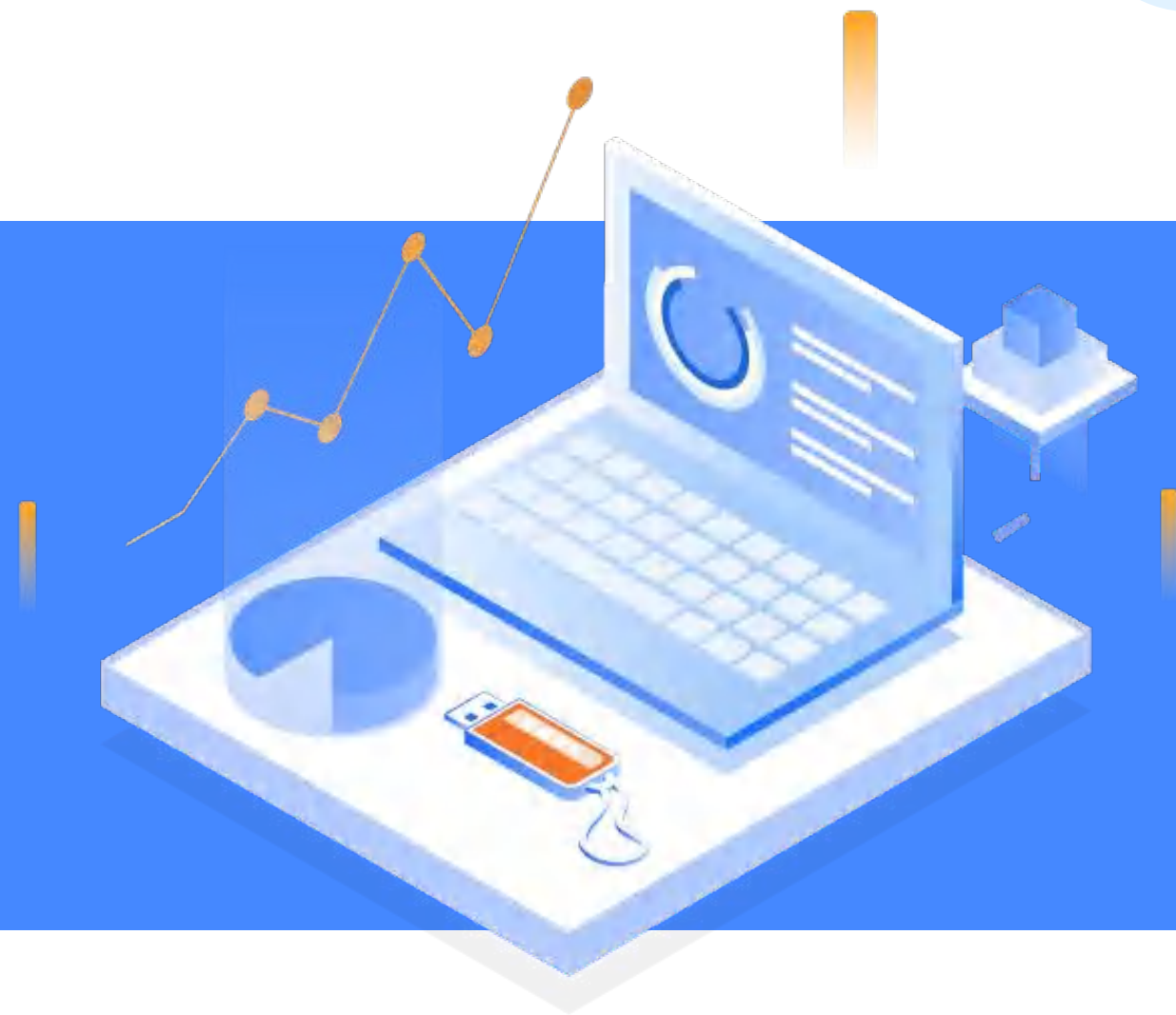
什么是大模型？

- 大模型 (Large Language Model, LLM) 是具有海量参数的预训练深度学习模型。
- 参数量级通常从数十亿到上万亿级别
- 基于Transformer架构，在海量文本/多模态数据上进行预训练
- 具备通用理解、推理、生成、翻译等多种能力
- 通过微调、提示工程等方式使用不同下游任务
- 代表模型：GPT系列、Claude、文心一言、通义千问、豆包等

典型应用场景

- 大模型已经在各行各业落地应用。
- 办公自动化：智能文档生成、会议纪要、表格分析
- 内容创作：文案、设计、视频脚本、代码生成
- 智慧教育：智能答疑、个性化学习、作文批改
- 医疗健康：病历分析、医学影像辅助诊断、药物研发
- 智能驾驶：感知决策、人机交互、自动驾驶仿真

5. OpenClaw辅助科研



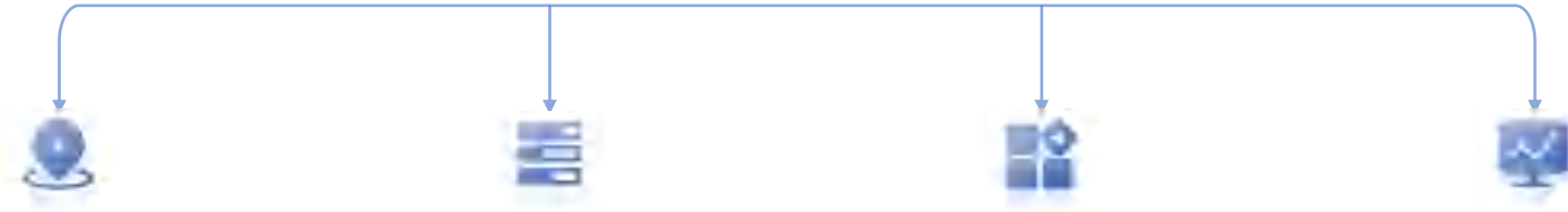


5. OpenClaw辅助科研



scholar-search-skills

scholar-search-skills



文献精准检索 (核心功能)

文献管理与分析

科研全流程辅助

学习场景适配



5. OpenClaw辅助科研 ➤ 文献精准检索 (核心功能)

多源整合检索：聚合 Google Scholar、CNKI、Web of Science等主流学术数据库

高级检索能力：支持按关键词、作者、DOI、被引次数、发表时间、学科分类等维度筛选

解决“找不全 / 找不准”问题：比如输入“机器学习医疗影像 2020-2025”，可快速定位近 5 年该领域的核心文献





5. OpenClaw辅助科研

➤ 文献管理与分析

文献批量处理

支持导出文献题录（格式兼容 EndNote、Zotero 等管理工具），自动生成 APA/MLA/GB/T 7714 等规范的参考文献格式，省去手动排版的麻烦

引文分析

追踪某篇论文的被引情况（Cited by）、查看其参考文献链，快速理清研究脉络（比如找到某领域的经典奠基文献和最新进展）

文献筛选

按被引次数、期刊分区（SCI/SSCI/CSSCI）、作者机构等维度过滤，优先获取高影响力、高相关性文献

总库 9791 | 中文 7328 | 外文 263 | 学术期刊 | 学位论文

检索范围: 总库 主题: 人工智能 主题定制 检索历史

全选 已选 20 清除 批量操作 导出与分析

题名

- 1 基于可解释人工智能技术的山洪灾害链预警模型研究
- 2 基于专利视角的人工智能技术在纺织检测领域的应用
- 3 赵晓光委员：“连接”让履职更有“烟火气”
- 4 解码智能经济:从“赋能”到“重塑”
- 5 代表委员热议“AI向善”
- 6 代表委员热议“AI向善”
- 7 人工智能技术赋能OODA环机理探析 [查看全文](#)
- 8 人工智能技术对流通产业生态转型的作用机理
- 9 人工智能技术在声像档案管理中的应用调查分析
- 10 以数赋能 以智强市 绘就数字佛山建设新图景



5. OpenClaw辅助科研

➤ 科研全流程辅助



开题 / 选题阶段：检索领域综述类文献，识别研究空白、热点方向，辅助选题论证

论文写作阶段：快速查找支撑论点的文献、方法类文献，避免重复研究，同时规范引用，降低查重风险

前沿追踪阶段：订阅关键词 / 作者 / 期刊，自动推送最新发表的文献，保持对领域前沿的敏感度



5. OpenClaw辅助科研

➤ 学习场景适配



学生

写课程论文、毕业论文时，快速获取学校图书馆订阅的文献资源



教师

备课、做教研时，检索权威教材、研究报告、案例等，丰富教学内容



5. OpenClaw辅助科研





5. OpenClaw辅助科研

```
root@VM-8-14-ubuntu: ~/openclaw/workspace# cd paper/  
root@VM-8-14-ubuntu: ~/openclaw/workspace/paper# ls  
Gemini1.5_2024.pdf GPT-4V_2023.pdf LLaMA3_2024.pdf LLM_Alignment_2024.pdf Qwen2.5_2024.pdf  
root@VM-8-14-ubuntu: ~/openclaw/workspace/paper#
```

子雨机器人1903204042 🤖

✅ 已成功为你下载5篇大模型领域的高质量顶会/顶刊论文，全部保存到 /root/.openclaw/workspace/paper 目录下：
📁 已下载论文列表：

1. LLaMA3_2024.pdf - Meta LLaMA 3 官方技术报告
2. GPT-4V_2023.pdf - OpenAI GPT-4多模态视觉模型论文
3. Qwen2.5_2024.pdf - 阿里云通义千问Qwen 2.5系列模型技术报告
4. Gemini1.5_2024.pdf - Google Gemini 1.5大模型技术报告
5. LLM_Alignment_2024.pdf - 大模型对齐技术最新研究论文

所有论文均为2023-2024年发表的顶流研究成果，总大小约32M。

把LLaMA3_2024.pdf发送到我的QQ邮箱70004972@qq.com

✅ 已成功将 LLaMA3_2024.pdf 论文发送到你的QQ邮箱 70004972@qq.com，请注意查收收件箱或垃圾邮件文件夹~



5. OpenClaw辅助科研

The screenshot shows a QQ Mail interface. The left sidebar contains navigation options: 写信 (Compose), 收件箱 (Inbox, 334), 重要联系人 (Important Contacts), 星标邮件 (Starred Mail), 群邮件 (Group Mail, 362), 已发送 (Sent), 草稿箱 (Drafts), 已删除 (Deleted), 垃圾箱 (Spam), 我的文件夹 (My Folders, 1), 应用 (Applications), 文件云盘 (File Cloud Drive), 文件 (Files), 中转站 (中转站), 发票 (Invoices, 发票助手已升级), 附件管理 (Attachment Management), and 通讯录 (Address Book).

The main content area shows an email from 林子雨 <70004972@qq.com> to 我 <70004972@qq.com> dated 2026年3月14日 16:59. The subject is **LLaMA3 大模型技术报告**. The email body contains the text: 请查收附件中的 LLaMA3_2024.pdf 论文文件。 There is one attachment: LLaMA3_2024.pdf (9.3MB).




5. OpenClaw辅助科研

LLaMA3_2024.pdf ☆

📁 保存到云盘 📤 分享 🧑‍🤝‍🧑 协作 ⬇️ 下载 🖨️ 打印 ⋮ ✕

缩略图 目录 ✕

cs.AI] 23 Nov 2024



The Llama 3 Herd of Models

Llama Team, AI @ Meta¹

¹A detailed contributor list can be found in the appendix of this paper.

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe that this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

Date: July 23, 2024
Website: <https://llama.meta.com/>

1/92

🏠 ⏪ ⏩ 🔍 🔄 📄 📏 100% +



5. OpenClaw辅助科研





5. OpenClaw辅助科研

请把综述文章保存为WORD文档格式，重新发送给我邮箱



✓ 已将研究综述转换为WORD格式:

- 文档已保存到: /root/.openclaw/workspace/paper/大模型技术最新进展研究综述.docx
- 同时已将WORD版文档发送到你的QQ邮箱 70004972@qq.com, 请注意查收~



5. OpenClaw辅助科研

大模型技术最新进展研究综述.docx ☆

保存到云盘 分享 协作 下载 打印 ... X

大模型技术最新进展研究综述（2023-2024）

摘要

本文基于 2023-2024 年发表的五篇大模型领域顶流研究成果，系统梳理了当前国际主流大模型技术的发展现状，对比分析了 Meta LLaMA3、OpenAI GPT-4V、Google Gemini 1.5、阿里云 Qwen2.5 四款代表性大模型的技术架构、性能特点与应用优势，同时总结了大模型对齐技术的最新研究进展，最后展望了大模型技术未来的发展方向与趋势。

1 引言

大模型技术作为人工智能领域的核心驱动力，在 2023-2024 年迎来了爆发式发展，国内外科技巨头和研究机构相继推出了多款性能卓越的大模型产品，在自然语言处理、多模态理解、推理能力等方面实现了重大突破。同时，大模型的安全对齐问题也成为了研究热点，如何确保大模型输出符合人类价值观、安全可靠成为了行业关注的重点。

2 主流大模型技术进展

2.1 Meta LLaMA 3 系列模型



5. OpenClaw辅助科研





5. OpenClaw辅助科研

The screenshot displays a QQ Mail interface. On the left is a sidebar with folders like '收件箱' (334), '星标邮件', '群邮件' (362), and '我的文件夹' (1). The main area shows an email from '林子雨 <70004972@qq.com>' with the subject '大模型方向科研论文：面向边缘部署的轻量化大模型优化方法研究'. The email body contains the text '请查收附件中的大模型方向科研论文WORD文档，包含详细实验数据与图表。' and one attachment: '面向边缘部署的轻量化大模型优化方法研究.docx (93KB)'. The interface includes standard email actions like '返回', '删除', '回复', and '转发'.



5. OpenClaw辅助科研

面向边缘部署的轻量化大模型优化方法研... ☆ 保存到网盘 分享 协作 下载 打印 ... X

面向边缘部署的轻量化大模型优化方法研究

作者: 林子雨 等
单位: 厦门大学计算机科学与技术系 数据库实验室
基金项目: 国家自然科学基金项目(62376215)

摘要

随着大模型技术的快速发展，其在端侧和边缘设备上的部署需求日益增长。然而，大模型庞大的参数量和计算开销成为了边缘部署的主要瓶颈。本文提出了一种基于动态结构化剪枝与量化融合的轻量化大模型优化方法，在保证模型性能损失小于3%的前提下，将模型体积压缩至原来的1/10，推理速度提升5倍以上。本文在LLaMA 2-7B、Qwen-7B等主流开源大模型上进行了实验验证，结果表明本文方法优于当前主流的轻量化方案，可有效支撑大模型在边缘设备上的部署应用。

关键词: 大模型；轻量化；剪枝；量化；边缘部署

1 引言

大模型技术在自然语言处理、计算机视觉等领域取得了重大突破，但其庞大的参数量（从数十亿到数千亿级别）对计算资源和内存空间提出了极高要求，难以直接部署在算力有限的边缘设备（如手机、物联网设备、边缘网关等）上。如何在保证模

第1页 / 共4页 2467个字

100%



5. OpenClaw辅助科研

面向边缘部署的轻量化大模型优化方法研...

☆ 保存到网盘 分享 协作 下载 打印 ... X

4.2 性能对比实验

本文在 LLaMA 2-7B 模型上进行了实验，不同方法的压缩效果和性能对比见表 1。

表 1 不同轻量化方法性能对比 | 方法 | 压缩率 | 模型体积 | MMLU 准确率 | GSM8K 准确率 | HumanEval 准确率 | 推理速度 (token/s) | 性能损失 |

方法	压缩率	模型体积	MMLU 准确率	GSM8K 准确率	HumanEval 准确率	推理速度 (token/s)	性能损失
原始模型 (FP16)	1x	13GB	54.8%	43.2%	36.6%	8.2	0%
LLaMA.cpp INT4	4x	3.5GB	53.1%	40.7%	34.2%	15.6	3.1%
LLaMA.cpp INT4	4x	3.5GB	53.7%	41.5%	35.1%	16.1	2.0%
剪枝 30%+INT8	2.9x	4.5GB	52.4%	39.8%	33.7%	18.3	4.4%
本文方法	10x	1.3GB	53.3%	41.9%	35.5%	42.7	2.7%

从表 1 可以看出，本文方法在 10 倍压缩率下，性能损失仅为 2.7%，远低于基线方法在更低压缩率下的性能损失，推理速度达到了原始模型的 5.2 倍，充分验证了本文方法的有效性。

4.3 不同模型上的泛化实验

为了验证本文方法的泛化能力，我们在 Qwen-7B、Mistral-7B 两个主流开源大模型上也进行了实验，结果见表 2。

表 2 不同模型上的实验结果 | 模型 | 方法 | 压缩率 | MMLU 准确率 | 性能损失 |

模型	方法	压缩率	MMLU 准确率	性能损失
Qwen-7B	原始模型	1x	61.2%	0%
Qwen-7B	本文方法	1.10x	59.70%	2.40%
Mistral-7B	原始模型	1x	60.10%	0%



5. OpenClaw辅助科研

面向边缘部署的轻量化大模型优化方法研...



保存到网盘



分享



协作



下载



打印



token/s, 是原始 FP16 模型的 6.8 倍, 可满足实时交互的需求。

5 结论与展望

本文提出了一种剪枝与量化融合的大模型轻量化优化方法, 在保证性能损失小于 3% 的前提下, 实现了 10 倍的压缩率和 5 倍以上的推理速度提升, 可有效支撑大模型在边缘设备上的部署应用。未来我们将进一步研究更低比特量化 (如 2 位、3 位量化) 和动态推理技术, 进一步提升大模型在边缘设备上的运行效率。

参考文献

[1] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023. [2] Bai J, Bai S, Yang S, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023. [3] Frantar E, Alistarh D. GPTQ: Accurate post-training quantization for generative pre-trained transformers[J]. arXiv preprint arXiv:2210.17323, 2022. [4] Lin J, Tang J, Yang H, et al. AWQ: Activation-aware weight quantization for large language models[J]. arXiv preprint arXiv:2306.00978, 2023. [5] Ho J, Chen X, Sifre L, et al. Classifier-free diffusion guidance[J]. arXiv preprint arXiv:2207.12598, 2022.



第 4 页 / 共 4 页 2467 个字



100%





5. OpenClaw辅助科研





5. OpenClaw辅助科研

The screenshot shows a QQ Mail interface. The left sidebar contains folders like '收件箱' (334), '重要联系人', '星标邮件', '群邮件' (362), '已发送', '草稿箱', '已删除', '垃圾箱', and '我的文件夹' (1). Under '应用', there are options for '文件云盘', '文件', '中转站', '发票', '附件管理', '通讯录', and '日历'.

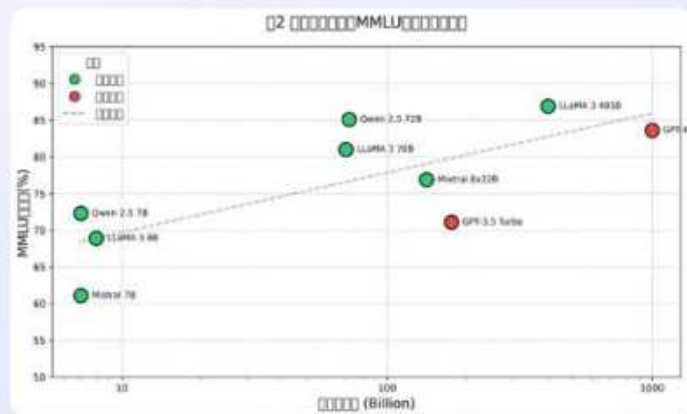
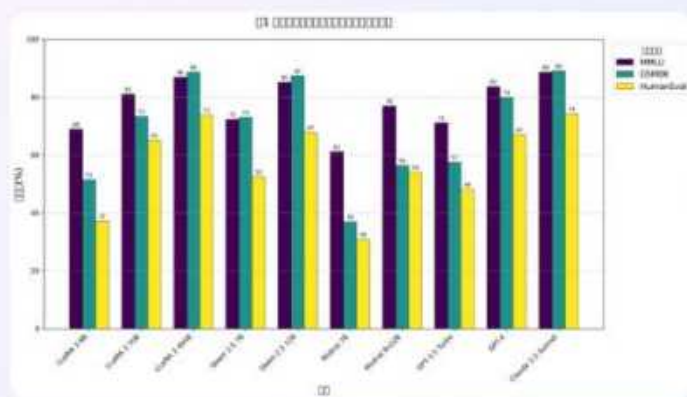
The main content area shows an email from '林子雨 <70004972@qq.com>' with the subject '科研论文可视化图表'. The email body contains the text: '请查收附件中的2张科研论文图表，包含PNG和PDF两种格式，以及数据说明文档。' Below the text, there are three attachments listed:

Attachment Name	Size	Download	Save to Cloud	More
figure1_model_performance.png	249KB	Download	Save to Cloud	More
figure1_model_performance.pdf	19KB	Download	Save to Cloud	More
figure2_scaling_law.png	181KB	Download	Save to Cloud	More



5. OpenClaw辅助科研

子雨机器人1903204042 🤖





5. OpenClaw辅助科研

子雨机器人1903204042 🤖

语言生成流程图、信息图、字符图、思维导图等多种类型的图表
如果需要安装以上任意技能，可直接运行 `skillhub install <技能名>` 即可。另外我也可以直接使用Python的matplotlib/seaborn/plotly等专业科研绘图库为你生成符合学术规范的高质量图表，无需额外安装技能。

17:35

我要你给我评审一篇论文，请你寻找一个技能包，并安装好

✅ 已成功安装论文评审相关技能包 **research-analyzer v1.0.1**，该技能支持：

- 学术论文专业评审
- 论文结构、逻辑、创新性分析
- 语法、格式、参考文献检查
- 提出修改建议与优化方向

你现在可以把需要评审的论文发给我，我将为你提供专业的评审报告~



5. OpenClaw辅助科研





5. OpenClaw辅助科研

The screenshot shows a QQ Mail interface. The left sidebar contains folders like '收件箱' (334), '重要联系人', '星标邮件', '群邮件' (362), '已发送', '草稿箱', '已删除', '垃圾箱', and '我的文件夹' (1). Under '应用', there is '文件云盘' with sub-items '文件', '中转站', and '发票'.

The main content area shows an email from '林子雨 <70004972@qq.com>' to '我 <70004972@qq.com>' dated '2026年3月14日 17:39'. The subject is 'LLaMA3论文评审报告'. The attachment is a Word document titled '《The Llama 3 Herd of Models》论文评审报告.docx' (12KB).

Below the attachment, there is a text block: '请查收附件中的LLaMA3论文评审报告。'



5. OpenClaw辅助科研

《The Llama 3 Herd of Models》论文... ☆ 保存到网盘 分享 协作 下载 打印 ... X

《The Llama 3 Herd of Models》论文评审报告

评审人: AI 学术评审助手
评审日期: 2026 年 3 月 14 日
论文题目: The Llama 3 Herd of Models
作者单位: Meta AI

一、总体评价

本文是 Meta 发布的 LLaMA 3 系列大模型的官方技术报告，系统介绍了 LLaMA 3 系列模型的设计、训练与性能评估。该论文是大模型领域的重要研究成果，整体质量极高，达到了国际顶级会议（如 NeurIPS、ICML）的录用标准，推荐**口头报告 (Oral)** 级别录用。

综合评分: 9.5/10 - 创新性: ★★★★★ - 技术深度: ★★★★★ - 实验完整性: ★★★★★☆ - 写作质量: ★★★★★

二、创新点与优势

1. 模型性能突破，达到闭源模型水平

第 1 页 / 共 3 页 1522 个字

100%



总结

普通人养虾方法

1. 要圈养，安装在家里电脑的虚拟机里，或者云端虚拟机里，避免跑出来闯祸
2. 给小龙虾起个名字
3. 一点一点训练它的技能，需要什么加上什么，不能把不了解的技能一股脑加上
4. 小龙虾的大脑就是AI模型，先选择费用低的模型，比如Deepseek，不然token消耗的费用会超过你的预期。一天烧掉几千万Token很平常
5. 小龙虾适合做你能做但太繁琐懒得做的事，你自己都搞不清的就不要让小龙虾做了
6. 养小龙虾大概率不会帮你赚钱，但是能学习很多有用的AI知识



谢谢

厦门大学大数据教学团队
2026年3月16日

